Reports

High-throughput sequencing of core STR loci for forensic genetic investigations using the Roche Genome Sequencer FLX platform

Sarah L. Fordyce[1,*], Maria C. Ávila-Arcos[1,*], Eszter Rockenbauer[2], Claus Børsting[2], Rune Frank-Hansen[2], Frederik Torp Petersen[2], Eske Willerslev[1], Anders J. Hansen[2], Niels Morling[2], and M. Thomas. P. Gilbert[1]
[1]*Centre for GeoGenetics, Natural History Museum of Denmark, Copenhagen, Denmark and* [2]*Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark*

BioTechniques 51:127-133 (August 2011) doi 10.2144/000113721
Keywords: Short tandem repeat (STR); high-throughput; sequencing; GS FLX

Supplementary material for this article is available at www.BioTechniques.com/article/113721.
*S.L.F. and M.C.A.-A. contributed to this work equally.

The analysis and profiling of short tandem repeat (STR) loci is routinely used in forensic genetics. Current methods to investigate STR loci, including PCR-based standard fragment analyses and capillary electrophoresis, only provide amplicon lengths that are used to estimate the number of STR repeat units. These methods do not allow for the full resolution of STR base composition that sequencing approaches could provide. Here we present an STR profiling method based on the use of the Roche Genome Sequencer (GS) FLX to simultaneously sequence multiple core STR loci. Using this method in combination with a bioinformatic tool designed specifically to analyze sequence lengths and frequencies, we found that GS FLX STR sequence data are comparable to conventional capillary electrophoresis-based STR typing. Furthermore, we found DNA base substitutions and repeat sequence variations that would not have been identified using conventional STR typing.

The analysis of short tandem repeat (STR) loci is an important tool in forensic genetics. Currently, STRs are mostly typed by DNA fragment analysis using PCR-based forensic STR kits and multicolor fluorescence capillary electrophoresis (CE) (1). Generally, the majority of STRs routinely investigated in forensic genetics are correlated to a well-defined allelic ladder repeat number. However, occasional variant or "off-ladder" alleles are observed (2–6). Current methods, although fast and cost-effective, do not allow full resolution of the STR loci. Recently, researchers have used pyrosequencing (7) or electrospray ionization mass spectrometry (ESI-MS) (8,9) to identify variants within or near repeat regions. However, with the ESI-MS method, only the base pair composition is obtained, and although variations compared with a reference sequence can be detected, variant positions cannot be determined. While a pyrosequencing approach was used to analyze STR loci in a low-throughput application (7), this methodology does not allow for the deep sequence coverage provided by high-throughput sequencing.

Since its inception in 2005, so-called high-throughput sequencing has become a rapid and reliable tool for massively parallel sequencing (10). The Roche FLX Genome Sequencer (GS FLX) system (and its related model, the GS Junior; both from Roche Applied Science, Indianapolis, IN, USA) is a pyrosequencing platform that can sequence hundreds of samples in parallel through the use of sample-specific multiple identifier (MID) tags. High-quality sequence reads of 400 bp or longer are routinely obtained, and the use of emulsion-PCR allows sequencing of one single bead-bound fragment within its own microreactor. Thus, the entire sequence of an STR allele can be obtained in a single experiment without confusion from competing reactions (e.g., from the allele on the other chromosome). While the GS FLX has proven to have some difficulties in accurately sequencing homopolymers (10), to date there has been no investigation into the sequencing of repeat regions using this technology for human DNA profiling. Until now, high-throughput sequencing has only been used for STR studies involving ancient DNA from extinct moas and for microsatellite location for species identification (11–14). These studies, however, used de novo sequencing to locate STRs rather than investigating well-defined STR regions.

Here, we examined the feasibility of applying GS FLX sequencing to STR typing for the purpose of forensic genetic investigations by directly comparing sequence data generated on the GS FLX platform to conventional CE-based STR methods using a novel bioinformatic tool. Although a laboratory method already exists for the sequencing of STRs (11,12), the present study addresses two major challenges: (*i*) the development of a bioinformatic tool that not only permits sorting of large amounts of data to facilitate data handling, but also creates an output file containing sequence lengths and frequencies, and (*ii*) the determination of the value of high-

throughput sequencing for STR profiling when compared with current CE methods. We present the successful characterization of a total of 10 individuals for 5 STR loci, including alleles that do not correlate with allele categories represented in the allelic ladder and the conventional CE (referred to as off ladder alleles) and a bioinformatic tool that can be applied to the analysis of the alleles for determining both the STR alleles and the ratios of sequences within an STR locus.

Material and methods

Samples

We tested five STR loci (CSF1PO, D13S317, D21S11, D5S818, and TH01), CSF1PO being the dominant locus tested, using 10 human samples that had previously been typed with the AmpF*l*STR Identifiler PCR Amplification kit (Applied Biosystems, Foster City, CA, USA) (Table 1). These samples were chosen as they contained either off-ladder alleles, three alleles at one locus, or were collected from a parent to an individual with three alleles at one locus. The samples were derived from either whole anticoagulated blood or buccal swabs (Table 1), from which DNA was extracted, PCR-amplified, and sequenced on the GS FLX. The samples were chosen to investigate the reliability of GS FLX sequencing of STRs for both regular alleles and for off-ladder alleles.

Extraction

The sample derived from buccal swabs was DNA-extracted using the BioRobot EZ1 Workstation (Qiagen, Valencia, CA, USA) (15), while the anticoagulated blood samples were extracted using the Tecan method (16).

PCR amplification

All samples were singleplex PCR-amplified in 25-µL reaction volumes using AmpliTaq Gold (Applied Biosystems). Supplementary Table S1 shows the primer sets used for each STR. Each reaction contained 1× AmpliTaq Gold buffer (Applied Biosystems), 2.5 mM $MgCl_2$ (Applied Biosystems), 0.5 µM forward and 0.5 µM reverse primers (DNA Technology A/S, Aarhus, Denmark), 0.2 mM each dNTP, 0.2 µL AmpliTaq Gold, and 1 µL DNA. The PCR cycling conditions were as follows: 94°C for 10 min, 30 cycles of 94°C for 30 s, 60°C for 30 s, and 72°C for 30 s, and a final extension at 72°C for 10 min.

Capillary electrophoresis

The samples were typed for five autosomal STRs (CSF1PO, D13S317, D21S11, D5S818, and TH01) using the AmpF*l*STR Identifiler PCR Amplification kit (Applied Biosystems) (17). In a second experiment, the samples were amplified with the primer set used in the GS FLX sequencing (Supplementary Table S1). The forward primers carried the 6-Fam dye at the 5′ end (DNA Technology A/S, Risskov, Denmark). PCR amplification conditions were as stated above. The PCR products were visualized on an AB Prism 3130XL Genetic Analyzer (Applied Biosystems) (17). The results are shown in Supplementary Table S3. The allele calls determined by the AmpF*l*STR Identifiler PCR Amplification kit were used for comparison with the allele calls from the sequencing data. Fragment lengths and sequence ratios from the second electrophoresis were used for comparison with similar parameters of the sequencing data (Table 1 and Supplementary Table S3).

GS FLX sequencing

The samples were converted into GS FLX libraries using a GS FLX Standard LR70 Sequencing kit following the manufacturer's instructions [GS FLX Shotgun DNA Library Preparation Method Manual (December 2007)]. During the library build, the samples were each given a unique MID tag, allowing the 10 samples to be pooled and sequenced together on one-eighth of a GS Pico TiterPlate (70 × 75; Roche Applied Science), along with 46 other samples not related to this study (therefore representing approximately one-fortieth of a GS FLX run). Emulsion PCR (emPCR) and subsequent bead-enrichment and sequencing steps were followed from the manufacturer's instructions [GS FLX emPCR Method Manual (December, 2007) and GS FLX Sequencing Method Manual (December, 2007), respectively].

Data sorting

An algorithm (available in the Supplementary Material) was designed to process the GS FLX FASTA reads by following a sequence of sorting/filtering steps. The reads were first sorted by MID tag into the original 10 libraries. Each library was then processed independently by the algorithm in search of STRs by identifying the presence on each read of at least one locus primer, and if so, this was followed by the identification, at each end of the read, of prespecified STR-end patterns. The prespecified STR-end patterns can be located anywhere along the amplified region. For this study, however, we chose prespecified sequences flanking the repeats regions only. This allowed only sequences containing the whole STR to be included rather than shorter fragments that would require assembly. If these patterns were identified, then the read was trimmed to its ends allowing a more precise estimation of the STR length. Finally, alignments and length distribution tables were generated for each STR locus.

Results and discussion

A total of 6488 sequence reads were successfully matched to our stringent algorithm, allowing only sequences containing locus primers and STR-delimiting patterns. Each sample contained on average 160 reads, which provided enough coverage for the program to make reliable calls and distinguish true variants from sequencing errors. Furthermore, it allowed grouping of the alleles according to repeat region lengths, and it enabled the sequence ratios of the alleles and stutters (a minor PCR artifact usually one repeat unit shorter than the parent allele) to be counted. A table containing the original data from the program, along with sequence lengths and sequence ratios is found in Supplementary Table S2. Consensus sequences for each allele are found in Supplementary Figures S1–S5 and Supplementary Table S4. The CE readouts of the conventional STR typing can be seen in Supplementary Table S3.

The accuracy of the GS FLX data for establishing allele and stutter sequence ratios was determined by comparing with the peak heights of the CE results (Supplementary Table S3). The peak information from the CE results contains a value for the size of the peak, while the GS FLX processed data gives the ratios of the various sequences (the alleles, stutters, and artifacts). Although these values are not directly comparable, the values were "normalized" (i.e., given a value out of 100) allowing comparisons to be made between the ratios of the sequenced alleles based on the two types of data (see Table 1). A Chi-square test was performed using the GS FLX data as the observed value and the CE data as the expected value. *P* values were determined and reported in Table 1. Overall, the frequencies of the alleles from both the CE and GS FLX data were equivalent with 6 of the 10 data sets ($P > 0.05$). For the seven samples identified as containing major stutters in the CE results (see Table 1), stutters were also seen in the GS FLX sequences. The stutters were one repeat unit shorter than the parent alleles, as expected based on the theory behind stutter formation and previous reports of stutters (18). The stutter percentage in the GS FLX data was 7.1%, with the

Table 1. Comparison of GS FLX data to CE-STR data.

System	Material	Sample	CE peaks		GS FLX reads		*P*
			Allele	Peak height ratio (%)	Allele	Seq ratio (%)	
CSF1PO	Blood	1	St	6	St	5	0.02
			11	37	11	52	
			12	40	12	30	
			13	17	13	13	
CSF1PO	Blood	2	St	5	St	5	0.98
			10	46	10	48	
			St	5	St	5	
			12	44	12	42	
CSF1PO	Blood	3	St	5	St	5	0.12
			10	48	10	38	
			10.3	47	10.3	57	
CSF1PO	Blood	4	St	9	St	8	0.73
			11	91	11	92	
CSF1PO	Blood	5	St	5	St	5	1.00
			10	50	10	50	
			11	45	11	45	
D13S317	Blood	6	6	54	6	62	0.11
			11	46	11	38	
D21S11	Blood	7	27/27.2	52	27.1	43	0.07
			31.2	48	31.2	57	
D21S11	Blood	8	St	7	St	15	0.01
			28	56	28	50	
			33.2	37	33.2	35	
D5S818	Swab	9	St	1	St	8	0.00
			11	99	11a	41	
					11b	51	
TH01	Blood	10	9.3	54	9.3	68	0.00
			10.2/11	46	10.3	32	

The results were normalized (given a value out of 100) to make the results of the two techniques comparable (see Supplementary Tables S2 and S3 for nonnormalized number of sequence). St, stutter; Seq, sequence. Allele 11a refers to the allele corresponding to the reference sequence (GenBank AC008512), and 11b refers to the allele with the G-to-T substitution. The column on the right contains the *P* values generated using a Chi-square test with 1 degree of freedom when comparing the CE peak height ratios to the GS FLX sequence ratios.

highest percentage coming from sample 8 with 15%. The average stutter percentage seen in the GS FLX data are consistent with that observed by fragment analysis using the AmpF*l*STR Identifiler PCR Amplification kit.

The lowest allele sequence ratio in our data set was found in sample 1 with only 13% of the sequences correlating to allele 13. This individual had three alleles at the CSF1PO locus. Allele 13 was also weakly amplified by the AmpF*l*STR Identifiler PCR Amplification kit. No mutation was found that could explain the low amplification efficiency in the flanking regions of the CSF1PO locus, and it remains unknown why allele 13 was weakly amplified. Apart from this rare triallelic sample, 32% of allele TH01 10.3 from individual 10 (Table 1) was the lowest allele sequence ratio for any individual. Other sequences not regarded as stutters (given that stutter sequences were one repeat unit shorter than the parent alleles) or the alleles themselves were considered to be sequencing errors (see Supplementary Table S2) and were not included in Table 1. Sequences that were of the same length as the alleles were counted in the ratio readouts. These included sequence reads with substitutions caused by sequencing errors. The artifacts listed in Supplementary Table S2 represent sequences corresponding to alleles with the exception of a 1-bp insertion or deletion. Single base pair insertions or deletions often occur when sequencing regions containing homopolymers (10,19). The sequencing artifacts could also be due to slippage of the DNA polymerase during replication, which has been observed in homopolymer and STR regions (20). All of these sequences were much less frequent (see Supplementary Table S2) than the stutters and alleles, consistent with previously applied background noise thresholds in CE methods (18).

The GS FLX sequence lengths of the alleles all corresponded to the lengths designated by the CE peaks, but samples 6, 8, and 9 showed differences from the reference sequences (www.cstl.nist.gov/strbase) that were not identified by CE. Another three samples showed previously unreported sequence variations either in the repeat structure or in the flanking regions (Supplementary Table S4). Sample 6 had an A to T substitution immediately after the repeat sequence at position 71 of allele 11 of the D13S317 locus (see Supplementary Figure S2). This not only differed from the reference sequence, but also from the other allele with six repeats. Sample 8, sequenced for the D21S11 locus (Supplementary Figure S3), had a 2-bp insertion before the last TCTA repeat of allele 33.2 of locus D21S11, similar to the D21S11 allele 31.2 of sample 7. Sample 8 also contained a T to C substitution in position 90 of the alignment (Supplementary Table S4). The D21S11 locus has a complex repeat structure, and the sequence data reveal much more information than fragment analysis by CE (Supplementary Figure S3). For example, the first part of the repeat sequences of allele 27.1 and 31.2 of D21S11 differs in sequence $[(TCTA)_6(TCTG)_5$ and $(TCTA)_5(TCTG)_6$, respectively] but not in length. Sample 9, which appeared to be homozygotic for the locus D5S818 when looking at the CE results, was found to be heterozygotic at D5S818 when looking at the sequencing results. One G-to-T substitution in the flanking region to the D5S818 STR sequence at position 63 of the alignment (see Supplementary Figure S4) resulted in two different alleles with the same repeat length (allele 11). This sample also had a T-to-C substitution at position 3 in both alleles compared with the reference sequence. This substitution, although in the flanking region before the repeats, is important, since it may aid in differentiating between individuals with 11 repeats, although further population studies are required to support this observation. New repeat sequences that have not been reported previously (Supplementary Table S4) were found in allele 10.3 of CSFPO1 from sample 3 and in allele 10.3 of TH01 from sample 10. Furthermore, a deletion of 3 bp (TCG) immediately after the repeat sequence (Supplementary Figure S3) was detected in allele 27.1 of the D21S11 locus in sample 7.

We have demonstrated that it is possible to generate high-coverage STR libraries from high-quality forensic blood

and swab templates. Our approach using high-throughput sequencing and data analysis produced reliable results, comparable to the CE approach. Furthermore, these results indicate that deep sequencing of the STRs allows better resolution of the STRs than traditional CE methods and fragment analysis. For instance, it is known that STRs are highly polymorphic within populations (21), hence it is no surprise that although two individuals have the same number of repeats at a certain locus, there are often single base pair substitutions or repeat structure variations that go undetected on CE, as previously demonstrated by pyrosequencing (7) and mass spectrometry (8,9). Within the small data set shown in this paper, we have demonstrated that 6 of the 10 samples deviate from previously reported repeat sequences in one or more of the five loci investigated. However, large population studies are required to estimate the true sequence variation of STR loci and to evaluate what impact the determination of these variations may have on future forensic genetic analyses.

Given the amount of coverage we generated even after stringent data sorting, it should be possible to sequence at least 400 PCR-amplified loci in one GS FLX sequencing run. This would render high-throughput sequencing a rapid, reliable, and economic tool for mass STR sequencing for forensic (or other) purposes. Furthermore, the GS FLX and the current titanium chemistry for the GS FLX allow sequences of up to 400–500 bp (http://454.com/products/gs-flx-system/index.asp) to be sequenced in one read. Hence, this technology is favored over other competing platforms (e.g., Illumina, SOLiD), as repeat regions targeted for use in forensic casework are generally between 100 and 500 bp, allowing the whole repeat region to be sequenced in one read.

Acknowledgments

We would like to gratefully acknowledge Morten Rasmussen for his advice concerning sequencing and Rune Frank-Hansen for discussions regarding the project idea. We would also like to gratefully acknowledge Charlotte Hallenberg, Bo Thisted Simonsen, Hanna E. Hansen, and Claus Børsting for collecting samples. S.L.F. and M.C.A.-A. wrote the manuscript. E.R., C.B., A.J.H., N.M., and M.T.P.G. edited the manuscript. Project idea was conceived by M.T.P.G., A.J.H., and experiments were designed by S.L.F., M.T.P.G., E.R., E.W., C.B., F.T.P., and A.J.H. M.C.A.-A. sorted the data and designed the algorithm.

Competing interests

The authors declare no competing interests.

References

1. **Butler, J.M., E. Buel, F. Crivellente, and B.R. McCord.** 2004. Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis. Electrophoresis *25*:1397-1412.
2. **Dauber, E.M., G. Dorner, S. Wenda, E.M. Schwartz-Jungl, B. Glock, W. Bär, and W.R. Mayr.** 2008. Unusual FGA and D19S433 off-ladder alleles and other allelic variants at the STR loci D8S1132, vWA, D18S51 and ACTBP2 (SE33). Forensic Sci. Int. Genet. Suppl. Series *1*:109-111.
3. **Onofri, V., M. Pesaresi, F. Alessandrini, C. Turchi, and A. Tagliabracci.** 2008. D16S539 microvariant or D2S1338 off-ladder allele? A case report about a range overlapping between two loci. Forensic Sci. Int. Genet. Suppl. Series *1*:123-124.
4. **Dauber, E.M., E.M. Schwartz-Jungl, S. Wenda, G. Dorner, B. Glock, and W.R. Mayr.** 2009. Further allelic variation at the STR-loci ACTBP2 (SE33), D3S1358, D8S1132, D18S51 and D21S11. Forensic Sci. Int. Genet. Suppl. Series *2*:41-42.
5. **Morales-Valverde, A., S. Silva-De La Fuente, G. Nunez-Rivas, and M. Espinoza-Esquivel.** 2009. Characterisation of 12 new alleles in the STR system D18S51. Forensic Sci. Int. Genet. Suppl. Series *2*:43-44.
6. **Kline, M.C., C.R. Hill, A.M. Decker, and J.M. Butler.** 2010. STR sequence analysis for characterizing normal, variant, and null alleles. Forensic Sci. Int. Genet. *5*:329-332.
7. **Divne, A.M., H. Edlund, and M. Allen.** 2010. Forensic analysis of autosomal STR markers using Pyrosequencing. Forensic Sci. Int. Genet. *4*:122-129.
8. **Planz, J.V., B. Budowle, T. Hall, A.J. Eisenberg, K.A. Sannes-Lowery, and S.A. Hofstadler.** 2009. Enhancing resolution and statistical power by utilizing mass spectrometry for detection of SNPs within the short tandem repeats. Forensic Sci. Int. Genet. Suppl. Series *2*:529-531.
9. **Pitterl, F., K. Schmidt, G. Huber, B. Zimmermann, R. Delport, S. Amory, B. Ludes, H. Oberacher, and W. Parson.** 2010. Increasing the discrimination power of forensic STR testing by employing high-performance mass spectrometry, as illustrated in indigenous South African and Central Asian populations. Int. J. Legal Med. *124*:551-558.
10. **Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, et al.** 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*:376-380.
11. **Allentoft, M., S.C. Schuster, R. Holdaway, M. Hale, E. McLay, C. Oskam, M.T. Gilbert, P. Spencer, et al.** 2009. Identification of microsatellites from extinct moa species using high-throughput (454) sequence data. BioTechniques *46*:195-200.
12. **Santana, Q., M. Coetzee, E. Steenkamp, O. Mlonyeni, G. Hammond, M. Wingfield, and B. Wingfield.** 2009. Microsatellite discovery by deep sequencing of enriched genomic libraries. BioTechniques *46*:217-223.
13. **Abdelkrim, J., B.C. Robertson, J.L. Stanton, and N.J. Gemmell.** 2009. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. BioTechniques *46*:185-191.
14. **Allentoft, M.E., C. Oskam, J. Houston, M.L. Hale, M.T.P. Gilbert, M. Rasmussen, P. Spencer, C. Jacomb, et al.** 2011. Profiling the dead: generating microsatellite data from fossil bones of extinct megafauna—protocols, problems, and prospects. PLoS One *6*:e16670.
15. **Stangegaard, M., M. Jørgensen, A.J. Hansen, and N. Morling.** 2009. Automated extraction of DNA from reference samples from various types of biological materials on the Qiagen BioRobot EZ1 workstation. Forensic Sci. Int. Genet. Suppl. Series *2*:69-70.
16. **Stangegaard, M., T.G. Frøslev, R. Frank-Hansen, S. Laursen, M. Jørgensen, A.J. Hansen, and N. Morling.** 2009. Automated extraction of DNA and PCR setup using a Tecan Freedom EVO Ukendt element: liquid handler. Progress in Forensic Genetics 13. Proceedings from the 23rd International ISFG Congress. Forensic Sci. Int. Genet. Suppl. Series *1*:74-76.
17. **Børsting, C., J.J. Sanchez, and N. Morling.** 2007. Forensic genetic DNA typing with PCR-based methods, p. 123-142. *In* S. Hughes and A. Moody (Eds.) PCR (Methods Express Series). Chapter 8. Scion Publishing Ltd, Bloxham, UK.
18. **Butler, J.M.** 2009. Fundamentals of Forensic DNA Typing. Elsevier, Burlington, MA.
19. **Droege, M. and B. Hill.** 2008. The Genome Sequencer FLX System—longer reads, more applications, straight forward bioinformatics and more complete data sets. J. Biotechnol. *136*:3-10.
20. **Seo, S.B., B.S. Jang, A. Zhang, J.A. Yi, H.Y. Kim, S.H. Yoo, Y.S. Lee, and S.D. Lee.** 2010. Alterations of length heteroplasmy in mitochondrial DNA under various amplification conditions. J. Forensic Sci. *55*:719-722.
21. **Jarne, P. and P.J.L. Lagoda.** 1996. Microsatellites, from molecules to populations and back. Trends Ecol. Evol. *11*:424-429.

Received 15 April 2011; accepted 23 June 2011.

Address correspondence to M. Thomas P. Gilbert, 1 Centre for GeoGenetics, Natural History Museum of Denmark, Øster Voldgade 5-7, 1350 Copenhagen K, Denmark. e-mail: tgilbert@snm.ku.dk