Microarray-based STR genotyping using RecA-mediated ligation

David Herrmann¹, Emily Rose¹, Uwe Müller² and Robert Wagner¹,*

¹Gene Check Inc., Greeley, CO 80634 and ²UWILA International Consulting, Alachua, FL 32615, USA

Received May 26, 2010; Revised July 1, 2010; Accepted July 12, 2010

ABSTRACT

We describe a novel assay capable of accurately determining the length of short tandem repeat (STR) alleles. STR genotyping is achieved utilizing RecA-mediated ligation (RML), which combines the high fidelity of RecA-mediated homology searching with allele-specific ligation. RecA catalyzes the pairing of synthetic oligonucleotides with one strand of a double-stranded DNA target, in this case a PCR amplicon. Ligation occurs only when two adjacent oligonucleotides are base paired to the STR region without any overlap or gap. RecA activity is required to overcome the inherent difficulty of annealing repeated sequences in register. This assay is capable of determining STR genotypes of human samples, is easily adapted to high throughput or automated systems and can have widespread utility in diagnostic and forensic applications.

INTRODUCTION

Tandem repeat sequences are found in the genomes of all higher eukaryotes. Upon local melting, regions of tandem repeat sequences are frequently able to form relatively stable secondary structures (e.g. cruciforms and slippage structures) (1). Such structures may serve as targets for nucleases (2–5) or other DNA-specific enzymes (6,7) and be sources of polymerase errors during replication, leading to expansion or contraction of the repeat region. Consequently, regions of tandem repeats typically exhibit substantial genetic instability (8,9), leading to heritable polymorphisms, genetic predisposition to disease and overt disease (10). One class of short tandem repeats (STRs), the microsatellites, contains repeat units of 1–4 bases and occurs in human DNA approximately once every 20 000 nt (11). The highly polymorphic nature of microsatellites makes them ideal markers for genetic mapping of specific loci or for distinguishing individual genomes (i.e. identity and paternity testing). In addition, they are used in medical applications, including identification of fetal cells in maternal blood (12) and monitoring of allogeneic bone marrow transplants (13).

A select group of STRs with a base repeat unit of 4 nt is used for forensic identification. In 1997, the FBI announced the selection of 13 STR loci to constitute the core of the US national database, the Combined DNA Index System or CODIS (14). Since then various commercial kits have been developed for STR genotyping. Depending on the number of STRs used, the discrimination power of the kits can range between 1:410 (3 STRs; Promega, released 1993) and 1:1.8 × 10¹⁷ (16 STRs; Promega PowerPlex 2.1). The CODIS system has also been adopted in many other countries, albeit with small variations in the number and type of STRs to match their needs based on the size and ethnic composition of their population.

Since a multiplex polymerase chain reaction (PCR) is used to amplify the chromosomal regions of interest (15), STR profiles can be determined with very small amounts of DNA. However, determination of the exact number of repeat units for each STR in the resulting PCR products requires their separation and accurate sizing. Traditionally this has been achieved by slab gel electrophoresis (16) or capillary electrophoresis (17,18). Capillary electrophoresis can be performed in microfabricated channels (19–21) or capillary arrays (22,23). More recently methods utilizing mass spectrometry (24) and microarray technology (25,26) have been developed in an attempt to reduce the time for analysis and/or to increase throughput.

We have developed a method of STR genotyping, RecA-mediated ligation (RML) that does not require the use of gel electrophoresis and which can be performed in less than three hours. RML combines oligonucleotide ligation with RecA-mediated homology searching. In conventional oligonucleotide ligation assays (OLA), two oligonucleotides are annealed to a single-stranded test DNA such that the 5' end of one oligonucleotide

*To whom correspondence should be addressed. Tel: +1 970 472 9951; Fax: +1 970 472 9956; Email: rwagner@genecheck.com

Present address:
David Herrmann, Cargill Research Fort Collins, CO 80525, USA.

© The Author(s) 2010. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
anneals adjacent to the 3′ end of the second oligonucleotide. If the adjacent ends are properly base paired and there is no overlap or gap between them, DNA ligase can act to covalently join the two oligonucleotides. OLA has been employed in SNP genotyping (27,28).

RecA, a bacterial recombinase, has been identified and most thoroughly characterized in Escherichia coli as the key player in the process of genetic recombination. It plays a crucial role in the search for and recognition of sequence homology and in the initial strand exchange process (29), and can be employed to catalyze strand exchange in vitro. Recombination is initiated when multiple RecA molecules coat a stretch of single-stranded DNA to form what is known as a RecA filament. In the presence of ATP, this protein–DNA complex searches double-stranded DNA for sequences homologous to the single-stranded oligonucleotide in the RecA filament. When homology is located, a three-stranded (D-loop) structure is formed wherein the RecA filament DNA is paired with the complementary strand of the duplex.

The precision of RecA-mediated homology searching and sequence alignment combined with the fact that this process requires a considerable stretch of homology [at least 30 bases in vivo (30)] allows RML to be used for genotyping of STR sequences, provided that non-repeat flanking sequences are included to insure proper alignment.

RecA-mediated ligation of short tandem repeats (RML-STR) (Figure 1) utilizes a set of allele-specific oligonucleotides for each STR sequence, each of which contains a different number of repeat sequences in addition to a stretch of sequence outside the repeat region. RecA-mediated homology searching uses this flanking sequence to align the oligonucleotide properly in a D-loop on the target DNA. A universal oligonucleotide containing a fixed number of repeat units and a flanking sequence from the opposite side of the repeat region will ligate only to that allele-specific oligonucleotide which forms a perfect ligation pair with the universal oligonucleotide, i.e. a D-loop containing the universal oligonucleotide and an allele-specific oligonucleotide aligned on the target DNA without overlap or gap. This system has been used to genotype the TPOX (human thyroid peroxidase gene) locus accurately from a number of samples. In the experiments described here, allele-specific oligonucleotides contained a terminal ‘zip code’ sequence at the 3′ end to allow them to be hybridized to capture oligonucleotides in separate spots on a microarray, and universal oligonucleotides were labeled to allow specific detection of RML ligation products.

**MATERIALS AND METHODS**

**Oligonucleotides and chemicals**

STR sequence data were obtained from the FBI Core CODIS loci website: (http://www.cstl.nist.gov/div831/str-base/fbicore.htm). The TPOX STR, located at intron 10 of the human thyroid peroxidase gene, was chosen for this study. Oligonucleotides were obtained from Integrated DNA Technologies (Coralville, IA) and were PAGE purified. The universal oligonucleotides contained (from 3′ to 5′) five tetranucleotide repeat units, 40 bases of sequence complementary to the sequence 3′ of the STR

![Figure 1. Rec A-mediated ligation genotyping of short tandem repeats.](https://academic.oup.com/nar/article-abstract/38/17/e172/1033322)
sequence, and a 5’ biotin-label. Each repeat specific oligonucleotide contained a unique 25mer immobilization sequence at its 5’ end and a sequence complementary to the region on the 5’-side of the STR repeat, in addition to between 0 and 9 tetranucleotide repeats and a 5’ phosphate group. Capture oligonucleotides were complementary to the immobilization sequence on the repeat specific oligonucleotides and were 5’ amine-labeled to allow immobilization to microarray slides. Immobilization sequences were provided by Nanosphere Inc. All chemicals were from Sigma unless otherwise specified. Molecular biology grade water was purchased from Fisher and was used for all RML and PCR reactions. Deionized water was used to make all other buffers.

PCR

Random human DNA samples were amplified in 50 μl reactions, 10 μl 5× Mix#4 buffer (Gene Check Inc.), 0.5 mM dNTPs (Bioline), 50 ng each of TPOX forward and reverse primers, 0.25 units Taq polymerase (Gene Check Inc.) and 75-ng genomic DNA. PCR was performed in Techne TC-512 thermocyclers under the following conditions: (i) 94°C for 3 min, (ii) 10 cycles of 94°C for 30 s, 66°C for 25 s and 72°C for 30 s, (iii) 20 cycles of 90°C for 30 s, 66°C for 25 s and 72°C for 30 s (iv) 66°C for 5 min and (v) held at 4°C.

We have also shown successful genotyping using target amplified with the KAPA Blood PCR Kit (KAPA Biosystems, Woburn, MA). This enabled PCR using 1 μl of whole blood and allowed deletion of the DNA extraction step from the standard protocol.

RML

A 10 μl reaction containing 50 mM Tris pH 7.5, 5 mM DTT, 1 mM MgCl2, 5 mM spermidine-acetate, 6.6 mM creatine phosphate (USB), 0.4 units creatine kinase (Roche), 2.8 mM ATP, 4 nM each repeat specific oligonucleotide, 40 nM universal oligonucleotide and 3.4 mM RecAE38K (Gene Check Inc.) was incubated at 37°C for 15 min to coat the oligonucleotides with RecA. After addition of PCR-amplified TPOX sequences and 0.5 units of T4 ligase, incubation was continued at 37°C for 30 min before stopping the reaction with 10 μl of 0.5 M EDTA, 0.1% Triton X-100.

Microarray detection method

RML reactions were diluted 1:2 by volume with 1× TBST + 0.5% casein and added to individual subarrays on a microarray slide and incubated at 50°C for 30 min to allow hybridization of ligation products to capture oligonucleotides immobilized to the microarray surface. Following 1× TBST wash, streptavidin-HRP (2 μg/ml) was added and incubated at room temperature (RT) for 10 min. Slides were washed, followed by addition of tyramide-Cy3 (1:50 in amplification diluent, Perkin Elmer) and continued incubation at room temperature for 10 min. Slides were washed, centrifuged to dry, and scanned for Cy3 fluorescence (Perkin Elmer ScanExpress). Data presented are averages of duplicate array spots measured for median signal intensity minus background.

All genotypes were confirmed by sequencing.

RESULTS AND DISCUSSION

STR (TPOX) genotyping via RML-STR

TPOX regions were amplified from human genomic DNA samples using primers to produce 240 base amplicons from the most common allele (eight repeats). The amplicons were mixed with RML oligonucleotides—eight allele-specific oligonucleotides containing from 1 to 8 repeat units and a universal oligonucleotide containing five repeat units. In this system the allele-specific oligonucleotides are specific for STR alleles containing five more repeats than the allele-specific oligonucleotide, for example, the three repeat unit allele-specific oligonucleotide is specific for an eight repeat allele, i.e. the five repeat-containing universal oligonucleotide and the three repeat allele-specific oligonucleotide will pair perfectly with eight repeat-containing target DNA.

Results of TPOX genotyping via RML-STR are presented in Figure 2. Clearly, RML-STR effectively genotypes the TPOX STR on a microarray platform. Although signal intensities vary somewhat, which may be a reflection of sample quality and/or PCR efficiency, the signals all exceed background by at least a factor of two and genotype calls are unambiguous. The signals are RecA dependent (Figure 3), indicating that they are not generated by ligation of RML oligonucleotides annealed to single strands remaining in the amplicons.

Presumably all oligonucleotides in an RML-STR reaction have nearly equal probabilities of forming D-loops. However, D-loops, at least those formed by single or pairs of unligated oligonucleotides, must be unstable. If such D-loops were stable, incorrect pairs could be immobilized as a consequence of being in a single D-loop and could, thereby, generate signal independent of ligation.

RML accomplishes homology searching, pairing and ligation without denaturation of the target DNA. Using double-stranded DNA effectively reduces the complexity of the RML reaction, relative to hybridization-based assays, and likely increases the sensitivity and utility of the RML reaction. RML-STR has been directly compared with a single-cycle OLA, which requires denaturation of the target DNA in the presence of potential ligation partners, hybridization of the oligonucleotides to the target and subsequent ligation (Figure 4). Allele-specific signals are, as expected, obtained with RML-STR. Very low and non-specific signal is obtained with OLA. The highest OLA signals are, as expected, obtained with allele-specific oligonucleotides, suggesting that the universal oligonucleotide and the three repeat allele-specific oligonucleotide will pair perfectly with eight repeat-containing target DNA.
Earlier work with RML in solution generally established that longer oligonucleotides perform better in RML than shorter oligonucleotides (data not shown). This may be a consequence of homology searching improving with increasing length of homology, which clearly appears to be true for oligonucleotides of less than around 40 nt. However, it may also be that once a critical length for homology searching is achieved, increased oligonucleotide length simply increases the likelihood of RecA filament formation. The rate limiting step in RecA filament formation is the nucleation step, i.e. binding of the first RecA monomer to a single-stranded DNA molecule. Once nucleation occurs, filament formation proceeds rapidly in a 5'→3' direction. Therefore, it may be that longer oligonucleotides favor nucleation by providing a larger target and, assuming nucleation is random with respect to location on the oligonucleotide, they would also increase the likelihood of forming sufficient extent of filament to institute successful homology searching. We attempted to distinguish these possibilities by examining three different universal oligonucleotides using an in solution assay. We tested a 60-base universal oligonucleotide fully complementary to the target DNA and two 85-base universal oligonucleotides, one fully complementary to the target and one with 60 bases complementary to the target and a 25-base non-complementary ‘tail’. The fully homologous 85-base oligonucleotide outperformed both other universal oligonucleotides (Figure 5). This result suggests that even better RML-STR genotyping than that reported above may be possible with longer RML oligonucleotides (the universal oligonucleotide in the experiments with immobilized allele-specific oligonucleotides was only 50 bases long).

Universal oligonucleotide length

A series of universal oligonucleotides containing various numbers of repeat units was tested. We found that
genotyping was possible with universal oligonucleotides containing from 1 to 8 repeat units (Figure 6). It should be noted that, when using the single repeat universal oligonucleotide, it was not possible to detect the 12 repeat allele, as we had no 11 repeat allele-specific oligonucleotide in the mix. Universal oligonucleotides containing from three to five repeats appear to work almost equally well (the universal oligonucleotide used in most...
experiments reported here contained five repeats). We have no good explanation for the very poor signal obtained with the six repeat universal oligonucleotide—it may simply have been a bad oligonucleotide preparation.

**Maximum STR length**

Because the ligation partner oligonucleotides used in RML-STR must span the entire length of the repeat region in an STR, there must be an upper STR length limit for successful STR genotyping. In other words, if the repeat region is too long, RecA filaments will be able to align with it independent of the flanking sequence, which will make accurate genotyping impossible. Results obtained with an 11 repeat universal oligonucleotide suggest that the maximum repeat length that can be accurately genotyped may be around 80 nt (20 repeats in the case of TPOX). When the 11 repeat universal oligonucleotide was used to genotype a 9,11 heterozygote, the 11 repeat genotype was easily detected (Figure 7). (The nine repeat allele could not be detected since it is smaller than the universal oligonucleotide.) However, when the same universal oligonucleotide was used to genotype an 8,12 heterozygote, both the 11 and 12 allele-specific oligonucleotides could ligate to the universal oligonucleotide. It may be that, when there are 44 bases of repeat sequence available for RecA to use in homology searching, i.e. in the filament oligonucleotide, the RecA filament can align without regard to the flanking sequence. The fact that the 12 repeat specific oligonucleotide did not ligate to the universal oligonucleotide with the 9,11 homozygote suggests that 40 bases of repeat sequence are not sufficient to align without the flanking sequence.

It may be possible to genotype STRs with more repeats by increasing the ratio of flanking sequence to repeat sequence in either universal or allele-specific oligonucleotides. Current oligonucleotide manufacturing technology does not allow high-yield production of oligonucleotides longer than ~100 nt, although advances in technology may make such oligonucleotides practical and affordable. However, it may be that the intrinsic properties of the RecA protein create an absolute limit to the amount of repeat sequence that can be included in either the universal or allele-specific oligonucleotides before alignment can be determined by the repeat sequences rather than the flanking sequences. If so, it may still be possible to genotype STRs with longer repeat regions by using three oligonucleotides and two ligation events. In such a system, there would be two universal oligonucleotides: (i) a labeled oligonucleotide containing a relatively low number of repeats and sufficient flanking sequence to assure alignment and (ii) a ‘bridging’ oligonucleotide containing only repeat sequence. Allele-specific oligonucleotides used in this system would need to be designed such that they could be ligated to the labeled universal only via the bridging oligonucleotide. For example, a 30 repeat STR might be genotyped using a 15 repeat allele-specific oligonucleotide, a five repeat labeled universal oligonucleotide

![Figure 6. TPOX STR genotyping by RML with 5' biotin-labeled universal oligonucleotides containing 1–8 units of the tetranucleotide repeat sequence. Each reaction included one of the biotin-labeled universal oligonucleotides (all 60 nt in length) at 40 nM. Reaction conditions and microarray detection method were as described in ‘Materials and Methods’ section.](https://academic.oup.com/nar/article-abstract/38/17/e172/1033322)
and a 10 repeat bridging oligonucleotide. Target DNA with more repeats would produce a hybridization product with a gap somewhere between the three oligonucleotides and target DNA with fewer repeats would produce a hybridization product with an overlap of the oligonucleotides. Neither structure would allow the two ligation events necessary for genotyping.

**RML-STR assay speed**

In order for an STR genotyping method to meet the demonstrated market need for rapid sample turnaround, it must be simple, i.e. involve few steps, and avoid long assay or hybridization times. In order to shorten assay times for RML-STR from PCR amplicons, we have adopted the KAPA Blood PCR system (KAPA Biosystems, Woburn, MA), which allows small blood samples (1 µl in our hands) to be used directly in PCR, avoiding the need for a separate extraction step (Figure 8). Use of KAPA has allowed us to reduce total assay to 3 h (Figure 9).

**RML-STR using multiplexed PCR amplicon target**

For RML-STR to be useful for human forensic, parentage and identity testing requires its application to genotyping of multiple STR loci. Based on our current estimates of maximum repeat length that can be genotyped by RML-STR, it appears that there are at least six additional CODIS loci that have repeat lengths appropriate to successful RML-STR genotyping. They are CSF1PO, TH01, D5S818, D8S1179, D13S317 and D16S539, all of which have fewer than 20 repeat units in their longest allele. This panel will have a random match probability for the most common genotype of <1 in 1.25 × 10^5. All seven loci (TPOX + six other CODIS loci) were amplified together in a multiplexed PCR reaction. Using this PCR reaction as target for TPOX RML-STR provided accurate genotypes as shown in Figure 10. However, the conditions employed appear not to be optimal as signal levels were lower than when TPOX-only PCR was used as target.

**Immobilized allele-specific oligonucleotides in RML-STR**

As shown in Figure 10, some depression of RML-STR signal was observed when using multiplexed PCR generated amplicons as target. Although genotyping was clearly possible and we believe that these signals could be improved with additional optimization, there is concern that the complexity of an in solution multiplexed RML-STR reaction (including more than 130 RML oligonucleotides for the seven CODIS STR panel) might be difficult to manage. The results of experiments mimicking multiplexed RML-STR by adding to the reaction mix additional RML oligonucleotides not complementary to

---

**Figure 7.** Maximum STR length that can be genotyped by RML-STR. Each reaction included a 5’ biotin-labeled universal oligonucleotide with 11 repeat units (85 nt in length) at 40 nM. Reaction conditions and microarray detection method were as described in ‘Materials and Methods’ section.
Figure 8. Genotyping of TPOX STR locus by RML-STR using KAPA PCR as target. TPOX sequences were amplified using 1.0 μl of whole blood in the KAPA Biosystems Blood PCR Kit. After PCR, the reaction was centrifuged and 4 μl of supernatant was used as target in standard RML-STR reaction. Reaction conditions and microarray detection method are as described in ‘Materials and Methods’ section.

Figure 9. Timeline of RML-STR.

Figure 10. RML-STR using multiplexed PCR amplicon as target. TPOX STR sequence was amplified in a multiplexed PCR with six additional primer pairs, specific for the other CODIS loci listed in the text. Successful amplification of all loci was confirmed by PAGE. Four micro liter of PCR reaction was used as target. Reaction conditions and microarray detection method were as described in ‘Materials and Methods’ section.
the target sequence, do show signal depression (Figure 11). One way to reduce oligonucleotide complexity is to immobilize the allele-specific oligonucleotides to individual microarray spots. Such immobilization would require only that the universal oligonucleotides (one per STR) be in solution and effectively create spatially separate reactions for each ligation pair in the RML reaction. Preliminary attempts at performing RML with immobilized oligonucleotides have been successful (Figure 12).

In order for a STR genotyping assay to have widespread utility it must accurately genotype a sufficient number of STR sequences to allow statistically significant discrimination power. The ability of the system to use double-stranded DNA target sequences coupled with the high fidelity of RecA-mediated homology searching should allow RML-STR to be easily adapted to simultaneous genotyping multiple STR sequences. It may even be possible to genotype directly from genomic DNA. The only limit to multiplexing RML-STR assays should be the number repeat units in an STR region. Some alleles of STR regions contain incomplete repeat units or single base insertions or deletions in individual repeat units. With the possible exception of those variants for which two alleles are of identical length and differ only by a single nucleotide substitution, it should be possible to genotype such alleles with RML-STR simply by including an oligonucleotide that contains the variant sequence.

The RecA-based STR genotyping method described here is simple, robust and rapid, does not require gel electrophoresis and is readily amenable to automation. The system should have widespread utility in research, forensic and diagnostic applications.

**ACKNOWLEDGEMENTS**

The authors would like to thank Steven Middleton for help with the figures, Miroslav Radman for critical reading of the manuscript and the laboratory of Michael Cox for providing us with a plasmid for the manufacture of RecA (E38K).

---

**Figure 11.** Effect of added oligonucleotides on RML-STR. Added oligonucleotides were 5' phosphate 85mers and a 5' biotin 60mer with non-specific sequences and were added to be at equal concentration to the TPOX oligonucleotides. Reaction conditions and microarray detection method were as described in 'Materials and Methods' section.
FUNDING
National Institutes of Health Small Business Innovation Research grant number 1R43HG004679-01 from the National Human Genome Research Institute. Funding for open access charge: Gene Check, Inc.

Conflict of interest statement. None declared.

REFERENCES

Figure 12. Immobilized allele-specific oligonucleotides in RML-STR. 5’ Amine-labeled, allele-specific oligos were immobilized to the microarray surface. Universal oligo (80 nM) (5’ P, 3’ biolabeled) and 4.5 mM RecA were added to the subarray and incubated at 37 °C for 15 min to allow filament formation. After addition of PCR-amplified TPOX sequences and T4 ligase, incubation continued at 37 °C for 60 min. Reaction was stopped by washing slide with 0.5 M NaOH. Signal development using TSA-Cy3 was as described in ‘Materials and Methods’ section.