Innovation, Sciences et Développement économique Canada
Office de la Propriété Intellectuelle du Canada

Innovation, Science and Economic Development Canada
Canadian Intellectual Property Office

CA 2907484 C 2021/06/29

(11)(21) 2 907 484

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(86) Date de dépôt PCT/PCT Filing Date: 2013/03/13
(87) Date publication PCT/PCT Publication Date: 2014/09/18
(45) Date de délivrance/Issue Date: 2021/06/29
(85) Entrée phase nationale/National Entry: 2015/07/16
(86) N° demande PCT/PCT Application No.: US 2013/030867
(87) N° publication PCT/PCT Publication No.: 2014/142831

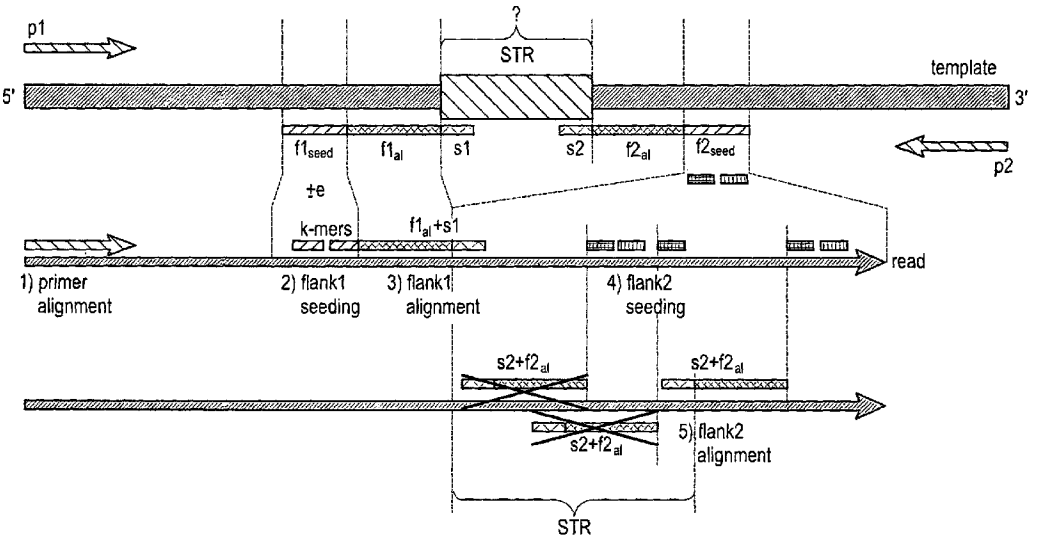
(51) Cl.Int./Int.Cl. *G16B 30/10* (2019.01), *C12Q 1/68* (2018.01), *C12Q 1/6809* (2018.01), *G16B 30/00* (2019.01)

(72) Inventeurs/Inventors:
BRUAND, JOCELYNE, US;
RICHARDSON, TOM, US;
MANN, TOBIAS, US

(73) Propriétaire/Owner:
ILLUMINA, INC., US

(74) Agent: GOWLING WLG (CANADA) LLP

(54) Titre : PROCEDES ET SYSTEMES POUR ALIGNER DES ELEMENTS D'ADN REPETITIFS
(54) Title: METHODS AND SYSTEMS FOR ALIGNING REPETITIVE DNA ELEMENTS




(57) Abrégé/Abstract:
Presented are methods and systems for aligning repetitive DNA elements. The methods and systems use the conserved flanks of repetitive polymorphic loci to effectively determine the length and sequence of the repetitive DNA element.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 September 2014 (18.09.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/142831 A1

(51) International Patent Classification:
*C12Q 1/68* (2006.01)

(21) International Application Number:
PCT/US2013/030867

(22) International Filing Date:
13 March 2013 (13.03.2013)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **ILLUMINA, INC.** [US/US]; 5200 Illumina Way, San Diego, California 92122 (US).

(72) Inventors: **BRUAND, Jocelyne**; 5200 Illumina Way, San Diego, California 92122 (US). **RICHARDSON, Tom**; 5200 Illumina Way, San Diego, California 92122 (US). **MANN, Tobias**; 5200 Illumina Way, San Diego, California 92122 (US).

(74) Agent: **MOORE, Brent C.**; 5200 Illumina Way, San Diego, CA 92122 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published**:

— *with international search report (Art. 21(3))*

(54) **Title**: METHODS AND SYSTEMS FOR ALIGNING REPETITIVE DNA ELEMENTS

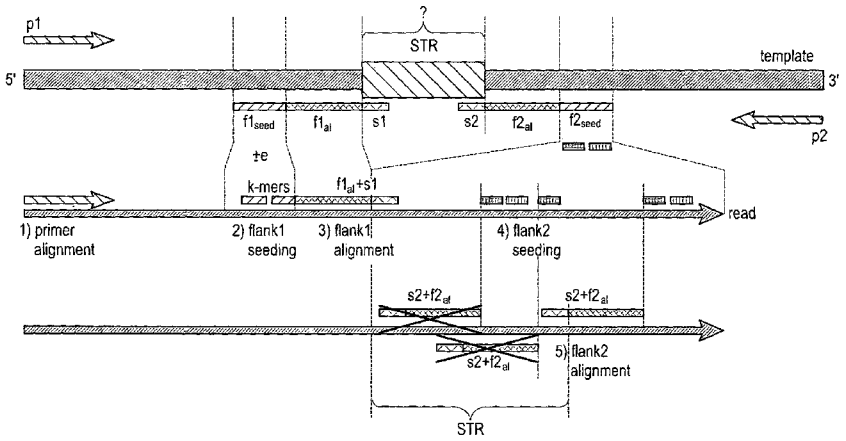



Fig. 1

(57) **Abstract**: Presented are methods and systems for aligning repetitive DNA elements. The methods and systems use the conserved flanks of repetitive polymorphic loci to effectively determine the length and sequence of the repetitive DNA element.

WO 2014/142831 A1

# METHODS AND SYSTEMS FOR ALIGNING REPETITIVE DNA ELEMENTS

## BACKGROUND

Sets of polymorphic, repetitive DNA elements are useful for many genetic applications including paternity testing, human identification (forensic DNA analysis), chimera monitoring (tissue transplantation monitoring), as well as many other uses in plant and animal genomics. One class of these repetitive elements comprises of the short tandem repeats (STRs). The allele of an STR locus is defined by its length, or number of repeat units, and by its sequence variation. While capillary electrophoresis systems can show the length of the allele, sequencing technologies have the additional differentiation power of discovering sequence variation, such as SNPs.

In order to take advantage of NGS data, it is advantageous to accurately and efficiently assign reads to the correct STR locus and STR allele. Existing methods for alignment of sequencing reads are time consuming and unable to detect all known and undiscovered polymorphic repetitive regions. As such, a great need exists for improved methods and systems for aligning repetitive DNA elements.

## BRIEF SUMMARY

Presented herein are methods and systems for aligning repetitive DNA elements. The methods and systems use the conserved flanks of repetitive polymorphic loci to effectively determine the length and sequence of the repetitive DNA element.

Accordingly, one embodiment presented herein is a method for determining the length of a polymorphic repetitive DNA element having a repeat region situated between a first conserved flanking region and a second conserved flanking region, the method comprising: (a) providing a data set comprising at least one sequence read of the polymorphic repetitive DNA element; (b) providing a reference sequence comprising the first conserved flanking region and the second conserved flanking region; (c) aligning a portion of the first flanking region of the reference sequence to the sequence read; (d) aligning a portion of the second flanking region of the reference sequence to the sequence read; and (e) determining the length and/or

sequence of the repeat region; wherein at least steps (c), (d) and (e) are performed using a suitably programmed computer. In certain embodiments, the aligning a portion of the flanking region in one or both of steps (c) and (d) comprises: (i) determining a location of a conserved flanking region on the read by using exact k-mer matching of a seeding region which overlaps or is adjacent to the repeat region; and (ii) aligning the flanking region to the sequence read. In some embodiments, the aligning can further comprise aligning both the flanking sequence and a short adjacent region comprising a portion of the repeat region.

Also presented herein is a system for determining the length of a polymorphic repetitive DNA element having a repeat region situated between a first conserved flanking region and a second conserved flanking region, the system comprising: a processor; and a program for determining the length of a polymorphic repetitive DNA element, the program comprising instructions for: (a) providing a data set comprising at least one sequence read of the polymorphic repetitive DNA element; (b) providing a reference sequence comprising the first conserved flanking region and the second conserved flanking region; (c) aligning a portion of the first flanking region of the reference sequence to the sequence read; (d) aligning a portion of the second flanking region of the reference sequence to the sequence read; and (e) determining the length and/or sequence of the repeat region; wherein at least steps (c), (d) and (e) are performed using a suitably programmed computer. In some embodiments, the aligning a portion of the flanking region in one or both of steps (c) and (d) comprises: (i) determining a location of a conserved flanking region on the read by using exact k-mer matching of a seeding region which overlaps or is adjacent to the repeat region; and (ii) aligning the flanking region to the sequence read. In some embodiments, the aligning can further comprise aligning both the flanking sequence and a short adjacent region comprising a portion of the repeat region.

In certain embodiments of the above methods or systems, the seeding region comprises a high-complexity region of the conserved flanking region, for example, the high-complexity region comprising sequence that is sufficiently distinct from the repeat region so as to avoid mis-alignment and/or a sequence having a diverse mixture of bases. In some embodiments, the seeding region avoids low-complexity regions of the conserved flanking region, for example sequence that substantially

resembles that of the repeat sequence and/or sequence having a mixture of bases with low diversity.

In certain embodiments of the above methods or systems, the seeding region is directly adjacent to the repeat region and/or comprises a portion of the repeat region. In certain embodiments, the seeding region is offset from the repeat region.

In certain embodiments of the above methods or systems, the dataset of sequence reads comprises sequence data from a PCR amplicon having a forward and reverse primer sequence. In certain embodiments, the at least one sequence read in the data set comprises a consensus sequence derived from multiple sequence reads. In certain embodiments, providing a reference sequence comprises identifying a locus of interest based upon the primer sequence of the PCR amplicon.

In certain embodiments of the above methods or systems, the repeat region is a short tandem repeat (STR) such as, for example, a STR selected from the CODIS autosomal STR loci, CODIS Y-STR loci, EU autosomal STR loci, EU Y-STR loci and the like.

Also presented herein is a method for determining the length and/or sequence of a polymorphic repetitive DNA element having a repeat region situated between a first conserved flanking region and a second conserved flanking region, the method comprising: (a) providing a data set comprising at least one sequence read of the polymorphic repetitive DNA element; (b) providing a reference sequence comprising the first conserved flanking region and the second conserved flanking region; (c) aligning a portion of the first flanking region of the reference sequence to the sequence read; (d) aligning a portion of the second flanking region of the reference sequence to the sequence read; and (e) determining the length and/or sequence of the repeat region; wherein at least steps (c), (d) and (e) are performed using a suitably programmed computer; wherein the aligning a portion of the flanking region in one or both of steps (c) and (d) comprises: (i) determining a location of a conserved flanking region on the read by using exact k-mer matching of a seeding region which overlaps or is adjacent to the repeat region; and (ii) aligning the flanking region to the sequence read; wherein the seeding region comprises a high-complexity region of the conserved flanking region, the high-complexity region comprising sequence that is sufficiently distinct from the repeat region so as to avoid mis-alignment.

Also presented herein is a system for determining the length and/or sequence of a polymorphic repetitive DNA element having a repeat region situated between a first conserved flanking region and a second conserved flanking region, the system comprising: a processor; and a program for determining the length and/or sequence of a polymorphic repetitive DNA element, the program comprising instructions for the processor to perform the following steps: (a) providing a data set comprising at least one sequence read of the polymorphic repetitive DNA element; (b) providing a reference sequence comprising the first conserved flanking region and the second conserved flanking region; (c) aligning a portion of the first flanking region of the reference sequence to the sequence read; (d) aligning a portion of the second flanking region of the reference sequence to the sequence read; and (e) determining the length and/or sequence of the repeat region; wherein the aligning a portion of the flanking region in one or both of steps (c) and (d) comprises: (i) determining a location of a conserved flanking region on the read by using exact k-mer matching of a seeding region which overlaps or is adjacent to the repeat region; and (ii) aligning the flanking region to the sequence read; wherein the seeding region comprises a high-complexity region of the conserved flanking region, the high-complexity region comprising sequence that is sufficiently distinct from the repeat region so as to avoid mis-alignment.

The details of one or more embodiments are set forth in the accompanying drawings and the description below. Other features, objects, and advantages will be apparent from the description and drawings, and from the claims.

## BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a schematic showing a method of alignment according to one embodiment.

Figure 2 is a schematic showing various mis-alignment errors that can occur if the flanking region immediately adjacent to the STR is used to seed the alignment.

Figure 3 is a set of graphs showing actual STR calling compared to theoretical results based on sample input from a mixture of samples.

Figure 4 is a table showing 100% concordance for allele calls for known loci of five control DNA samples.

## DETAILED DESCRIPTION

Sets of polymorphic, repetitive DNA elements are useful for many genetic applications including paternity testing, human identification (forensic DNA analysis), chimera monitoring (tissue transplantation monitoring), as well as many

other uses in plant and animal genomics. In order to take advantage of next generation sequencing (NGS) data, tools are needed for accurate and efficient assignment of sequencing reads to the correct repetitive DNA element locus and allele. One class of these repetitive elements comprises of the short tandem repeats (STRs). The allele of an STR locus is defined by its length, or number of repeat units, and by its sequence variation. While capillary electrophoresis systems can show the length of the allele, sequencing technologies have the additional differentiation power of discovering sequence variation, such as SNPs. It will be appreciated that although the methods and systems described herein are discussed in the context of STRs, they can be applied to any other repetitive DNA element.

Existing alignment methods fail for various reasons. One common approach is alignment to a reference sequence is commonly performed. However, the difference in allele sizes greatly differs, even within a single locus. For example, one core U.S. locus, FGA, has known alleles between 12.2 and 51.2, involving differences of 156 nucleotides (or even greater). Most aligners will not align reads with such a large gap, and any alleles which are too far from a reference sequence will be discarded by the aligner.

Another existing approach with drawbacks is the method of aligning to a reference ladder. Typically, a "reference genome" is created by building a ladder of all known STR alleles and aligning the reads to this reference, as typically done with NGS whole genome sequence data or targeted sequencing of non-repetitive DNA regions. There are shortcomings to this method. For example, known information about the STR sequence, such as primer sequence or conserved flanking regions, is ignored. Existing ladders are incomplete, since the sequences of many polymorphic repetitive regions are currently unknown. Due the highly variable nature of these genomic regions, new alleles may be discovered in the future. Further, changes to the sequence of one allele in the reference may have global effects to the reads alignment due to homology between the sequences.

Another alternative methodology for detecting STRs, known as lobSTR, senses then calls all existing STRs from sequencing data of a single sample *de novo*, with no prior knowledge of the STRs (*see* Gymrek et al. 2012 Genome Research 22:1154-62). However, the lobSTR method ignores prior knowledge (primer sequences, flanking regions) and miscalls some alleles. Further lobSTR misses STR

loci with complex repeat patterns, including some from the CODIS such as D21S11, allele 24 ($[TCTA]_4[TCTG]_6[TCTA]_3TA[TCTA]_3TCA[TCTA]_2TCCA\ TA[TCTA]_6$) or vWA, allele 16 ($TCTA[TCTG]_3[TCTA]_{12}TCCA\ TCTA$). Further, lobSTR assumes homozygous or heterozygous alleles, and is therefore not useful for handling samples having mixtures.

Thus, there exists a great need for a targeted approach utilizing prior knowledge greatly increases sensitivity and specificity.

Presented herein are methods and systems which use the conserved flanks of repetitive polymorphic loci to effectively determine the sequence of the repetitive DNA element. The methods advantageously align the beginning of the read sequence to the possible primer sequences to establish the locus and strand to which the read corresponds. Then, sections of the appropriate flanking sequences on each side of the repetitive locus are aligned to the read in order to pull the exact length and sequence from the read. These alignments are seeded using a k-mer strategy. The seed regions can be, for example, in a pre-chosen high-complexity region of the flanking sequence, close to the repeat region, but avoiding low-complexity sequence with homology to the target locus. This approach advantageously avoids misalignment of low-complexity flanking sequences close to the repeat region of interest.

The approach described herein is novel, and is surprisingly effective in properly determining the allele size and sequence. The methods make use of known sequences in the flanks of the STR themselves, which have been previously defined based on the known existing variations among the human population. Advantageously, performing alignment of a short span of flanking regions is computationally quick when compared to other methods. For example, a dynamic programming alignment (Smith-Waterman type) of the entire read is CPU intensive, time consuming, especially where multiple sequence reads are to be aligned. Furthermore, time spent aligning an entire sequence (for which a reference may not even exist) takes up valuable computational resources.

Using flanking regions to properly determine the allele provides several other unexpected advantages over existing methods. For example, BWA, a typical aligner, performs poorly when it is used to align to a reference, primarily due to the repetitive nature of an STR sequence and the incomplete state of the reference.

Further, the inventors have observed that changing the reference for one STR locus often affected calls for another locus, which should be independent. However, because forensics applications require high confidence calls, there is very little room for error.

Additional embodiments of the methods provided herein identify unique seeds within a flanking sequence. This approach allows for a reduction in alignment time and plays a role in avoiding misalignments in the case of low-complexity flanks.

The methods presented herein make use of prior knowledge of flanking sequence to ensure the proper call of the STR allele. In contrast, existing methods, which rely on a full reference sequence for each allele, face significant failure rates in situations where there is an incomplete reference. There are many alleles for which the sequence is not known, and possibly some yet unknown alleles. By way of illustration, assume a locus with a simple repeat pattern [TCTA] and a 3' flank starting with the sequence TCAGCTA. Thus, the reference may include such sequences as [flank1][TCTA]$_n$TCAGCTA[rest_of_flank2], where n is the number of repeats in the allele. The 9.3 allele would differ from the 10 allele by having a deletion somewhere along the sequence. Hopefully, these would be included in the reference, though it could be that not all are. [TCTA]$_7$TCA[TCTA]$_2$ is an example of such an allele. Under existing alignment protocols, any read ending after the [TCTA]$_7$ and before the final [TCTA], will align to [flank1][TCTA]$_7$TCAGCTA, making an improper call.

**Alignment Methods**

The methods provided herein allow for determining the length of a polymorphic repetitive DNA element having a repeat region situated between a first conserved flanking region and a second conserved flanking region. In one embodiment, the methods comprise providing a data set comprising at least one sequence read of a polymorphic repetitive DNA element; providing a reference sequence comprising the first conserved flanking region and the second conserved flanking region; aligning a portion of the first flanking region of the reference sequence to the sequence read; aligning a portion of the second flanking region of the reference sequence to the sequence read; and determining the length and/or

sequence of the repeat region. In typical embodiments, one or more steps in the method are performed using a suitably programmed computer.

As used herein, the term "sequence read" refers to sequence data for which the length and/or identity of the repetitive element are to be determined. The sequence read can comprise all of the repetitive element, or a portion thereof. The sequence read can further comprise a conserved flanking region on one end of the repetitive element (e.g., a 5' flanking region). The sequence read can further comprise an additional conserved flanking region on another end of the repetitive element (e.g., a 3' flanking region). In typical embodiments, the sequence read comprises sequence data from a PCR amplicon having a forward and reverse primer sequence. The sequence data can be obtained from any suitable sequence methodology. The sequencing read can be, for example, from a sequencing-by-synthesis (SBS) reaction, a sequencing-by-ligation reaction, or any other suitable sequencing methodology for which it is desired to determine the length and/or identity of a repetitive element. The sequence read can be a consensus sequence derived from multiple sequence reads. In certain embodiments, providing a reference sequence comprises identifying a locus of interest based upon the primer sequence of the PCR amplicon.

As used herein, the term "polymorphic repetitive DNA element" refers to any repeating DNA sequence, and the methods provided herein can be used to align the corresponding flanking regions of any such repeating DNA sequence. The methods presented herein can be used for any repeat region. The methods presented herein can be used for any region which is difficult to align, regardless of the repeat class. The method presented herein are especially useful for a region having conserved flanking regions. Additionally or alternatively, the methods presented herein are especially useful for sequencing reads which span the entire repeat region including at least a portion of each flanking region. In typical embodiments, the repetitive DNA element is a variable number tandem repeat (VNTR). VNTRs are polymorphisms where a particular sequence is repeated at that locus numerous times. Some VNTRs include minisatellites, and microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs). In some embodiments, the repetitive sequence is typically less than 20 base pairs, although larger repeating units can be aligned. For example, in typical embodiments, the

repeating unit can be 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more nucleotides, and can be repeated up to 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 or up to at least 100 times or more. In certain embodiments, the polymorphic repetitive DNA element is an STR. In some embodiments, the STR is used for forensic purposes. In typical embodiments for forensic applications, for example, the polymorphic repetitive DNA element comprises tetra- or penta-nucleotide repeat units, however, the methods provided herein are suitable for any length of repeating unit. In certain embodiments, the repeat region is a short tandem repeat (STR) such as, for example, a STR selected from the CODIS autosomal STR loci, CODIS Y-STR loci, EU autosomal STR loci, EU Y-STR loci and the like. As an example, the CODIS (Combined DNA Index System) database is a set of core STR loci for identified by the FBI laboratory and includes 13 loci: CSF1PO, FGA, TH01, TPOX, VWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51 and D21S11. Additional STRs of interest to the forensic community and which can be aligned using the methods and systems provided herein include PENTA D and PENTA E. The methods and systems presented herein can be applied to any repetitive DNA element and are not limited to the STRs described above. As used herein, the term "reference sequence" refers to a known sequence which acts as a scaffold against which a sample sequence can be aligned. In typical embodiments of the methods and systems provided herein, the reference sequence comprises at least a first conserved flanking region and a second conserved flanking region. The term "conserved flanking region" refers to a region of sequence outside the repeat region. The region is typically conserved across many alleles, even though the repeat region may be polymorphic. A conserved flanking region as used herein typically will be of higher complexity than the repeat region. In typical embodiments, a single reference sequence can be used to align all alleles within a locus. In some embodiments, more than one reference sequence is used to align all alleles within a locus because of variation within the flanking region. For example, the repeat region for Amelogenin has differences in the flanks between X and Y, although a single reference can represent the repeat region if a longer region is included in the reference.

In embodiments presented herein a portion of a flanking region of a reference sequence is aligned to the sequence read. Aligning is performed by determining a location of the conserved flanking region and then conducting a sequence alignment of that portion of the flanking region with the corresponding portion of the sequence read. Aligning of a portion of a flanking region is performed according to known alignment methods. In certain embodiments, the aligning a portion of the flanking region in one or both of steps (c) and (d) comprises: (i) determining a location of a conserved flanking region on the read by using exact k-mer matching of a seeding region which overlaps or is adjacent to the repeat region; and (ii) aligning the flanking region to the sequence read. In some embodiments, the aligning can further comprise aligning both the flanking sequence and a short adjacent region comprising a portion of the repeat region.

An example of this approach is illustrated in Figure 1. An amplicon ("template") is shown in Figure 1 having a STR of unknown length and/or identity. As shown in Figure 1, an initial primer alignment is conducted to identify the locus of interest, in this case an STR. The primers are illustrated as p1 and p2, which are the primer sequences that were used to generate the amplicon. In the embodiment shown in Figure 1, p1 alone is used during the primer alignment step. In some embodiments, p2 alone is used for primer alignment. In other embodiments, both p1 and p2 are used for primer alignment.

Following primer alignment, flank 1 is aligned, designated in Figure 1 as $fl_{al}$. Flank 1 alignment can be preceded by seeding of flank 1, designated in Figure 1 as $fl_{seed}$. Flank 1 seeding to correct for a small number (e) of indels between the beginning of the read and the STR. The seeding region may be directly next to the beginning of the STR, or may be offset (as in figure) to avoid low-complexity regions. Seeding can be done by exact k-mer matching.

Flank1 alignment proceeds to determine the beginning position of the STR sequence. If the STR pattern is conserved enough to predict the first few nucleotides (s1), these are added to the alignment for improved accuracy.

Since the length of the STR is unknown, an alignment is performed for flank2 as follows. Flank2 seeding is performed to quickly find out possible end positions of the STR. As the seeding for flank 1, the seeding may be offset to avoid low-complexity regions and mis-alignment. Any flank 2 seeds that fail to align are

discarded. Once flank2 properly aligns, the end position (s2) of the STR can be determined, and the length of the STR can be calculated.

The seeding region can directly adjacent to the repeat region and/or comprises a portion of the repeat region. In some embodiments, the location of the seeding region will depend on the complexity of the region directly adjacent to the repeat region. The beginning or end of an STR may be bounded by sequence that comprises additional repeats or which has low complexity. Thus, it can be advantageous to offset the seeding of the flanking region in order to avoid regions of low complexity. As used herein, the term "low-complexity" refers to a region with sequence that resembles that of the repeat sequence. Additionally or alternatively, a low-complexity region incorporates a low diversity of nucleotides. For example, in some embodiments, a low-complexity region comprises sequence having more than 30%, 40%, 50%, 60%, 70% or more than 80% sequence identity to the repeat sequence. In typical embodiments, the low-complexity region incorporates each of the four nucleotides at a frequency of less than 20%, 15%, 10% or less than 5% of all the nucleotides in the region. Any suitable method may be utilized to determine a region of low-complexity. Methods of determining a region of low-complexity are known in the art, as exemplified by the methods disclosed in Morgulis et al., (2006) Bioinformatics. 22(2):134-41. For example, as described in the incorporated materials for Morgulis et al., an algorithm such as DUST may be used to identify regions within a given nucleotide sequence that have low complexity.

In some embodiments, the seeding is offset from the start of the STR by at least 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40 or more nucleotides. In some embodiments, the flanking region is evaluated to identify a region of high complexity. As used herein, the term "high-complexity region" refers to a region with sequence that is different enough from that of repeat that it removes possibilities of mis-alignments. Additionally or alternatively, a high complexity region incorporates a variety of nucleotides. For example, in some embodiments, a high-complexity region comprises sequence having less than 80%, 70%, 60%, 50%, 40%, 30%, 20% or less than 10% identity to the repeat sequence. In typical embodiments, the high-complexity region incorporates each of the four nucleotides

at a frequency of at least 10%, 15%, 20%, or at least 25% of all the nucleotides in the region.

As used herein, the term "exact k-mer matching" refers to a method to find optimal alignment by using a word method where the word length is defined as having a value *k*. In some embodiments, the value of k is 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40 or more nucleotides in length. In typical embodiments, k has a value of between 5 and 30 nucleotides in length. In some typical embodiments, k has a value of between 5 and 16 nucleotides in length. In certain embodiments, k is chosen on-line. For example, if a flank region is short (primer close to the STR), k is reduced appropriately. In typical embodiments, k is chosen so as to guarantee finding all matches with edit distance e. Word methods identify a series of short, nonoverlapping subsequences ("words") in the query sequence that are then matched to candidate database sequences. The relative positions of the word in the two sequences being compared are subtracted to obtain an offset; this will indicate a region of alignment if multiple distinct words produce the same offset. Only if this region is detected do these methods apply more sensitive alignment criteria; thus, many unnecessary comparisons with sequences of no appreciable similarity are eliminated. Methods of performing k-mer matching, including exact k-mer matching, are well known in the art, as exemplified by the disclosure of Lipman, et al., (1985) *Science* 227:1435–41, and of Altschul, et al., (1990) *Journal of Molecular Biology* **215**:403–410.

In certain embodiments, providing a reference sequence comprises identifying a locus of interest based upon the primer sequence of an amplicon. As used herein, the term "amplicon" refers to any suitable amplification product for which is a sequence is obtained. Typically, the amplification product is a product of a selective amplification methodology, using target-specific primers, such as PCR primers. In certain embodiments, the sequence data is from a PCR amplicon having a forward and reverse primer sequence. In some embodiments, selectively amplifying can include one or more non-selective amplification steps. For example, an amplification process using random or degenerate primers can be followed by one or more cycles of amplification using target-specific primers. Suitable methods

for selective amplification include, but are not limited to, the polymerase chain reaction (PCR), strand displacement amplification (SDA), transcription mediated amplification (TMA) and nucleic acid sequence based amplification (NASBA), as described in U.S. Patent No. 8,003,354. The above amplification methods can be employed to selectively amplify one or more nucleic acids of interest. For example, PCR, including multiplex PCR, SDA, TMA, NASBA and the like can be utilized to selectively amplify one or more nucleic acids of interest. In such embodiments, primers directed specifically to the nucleic acid of interest are included in the amplification reaction. Other suitable methods for amplification of nucleic acids can include oligonucleotide extension and ligation, rolling circle amplification (RCA) (Lizardi et al., Nat. Genet. 19:225-232 (1998)) and oligonucleotide ligation assay (OLA) (See generally U.S. Pat. Nos. 7,582,420, 5,185,243, 5,679,524 and 5,573,907; EP 0 320 308 B1; EP 0 336 731 B1; EP 0 439 182 B1; WO 90/01069; WO 89/12696; and WO 89/09835) technologies. It will be appreciated that these amplification methodologies can be designed to selectively amplify a target nucleic acid of interest. For example, in some embodiments, the selective amplification method can include ligation probe amplification or oligonucleotide ligation assay (OLA) reactions that contain primers directed specifically to the nucleic acid of interest. In some embodiments, the selective amplification method can include a primer extension-ligation reaction that contains primers directed specifically to the nucleic acid of interest. As a non-limiting example of primer extension and ligation primers that can be specifically designed to amplify a nucleic acid of interest, the amplification can include primers used for the GoldenGate™ assay (Illumina, Inc., San Diego, CA), as described in U.S. Pat. No. 7,582,420. The present methods are not limited to any particular amplification technique and amplification techniques described herein are exemplary only with regard to methods and embodiments of the present disclosure.

Primers for amplification of a repetitive DNA element typically hybridize to the unique sequences of flanking regions. Primers can be designed and generated according to any suitable methodology. Design of primers for flanking regions of repeat regions is well known in the art, as exemplified by Zhi, et al. (2006)

*Genome Biol*, 7(1):R7. For example, primers can be designed manually. This involves searching the genomic DNA sequence for microsatellite repeats, which can be done by eye or by using automated tools such as RepeatMasker software. Once the repeat regions and the corresponding flanking regions are determined, the flanking sequences can be used to design oligonucleotide primers which will amplify the specific repeat in a PCR reaction.

Systems

Also presented herein is a system for determining the length of a polymorphic repetitive DNA element having a repeat region situated between a first conserved flanking region and a second conserved flanking region, the system comprising: a processor; and a program for determining the length of a polymorphic repetitive DNA element, the program comprising instructions for: (a) providing a data set comprising at least one sequence read of the polymorphic repetitive DNA element; (b) providing a reference sequence comprising the first conserved flanking region and the second conserved flanking region; (c) aligning a portion of the first flanking region of the reference sequence to the sequence read; (d) aligning a portion of the second flanking region of the reference sequence to the sequence read; and (e) determining the length and/or sequence of the repeat region; wherein at least steps (c), (d) and (e) are performed using a suitably programmed computer. In some embodiments, the aligning a portion of the flanking region in one or both of steps (c) and (d) comprises: (i) determining a location of a conserved flanking region on the read by using exact k-mer matching of a seeding region which overlaps or is adjacent to the repeat region; and (ii) aligning the flanking region to the sequence read. In some embodiments, the aligning can further comprise aligning both the flanking sequence and a short adjacent region comprising a portion of the repeat region.

A system capable of carrying out a method set forth herein can be, but need not be, integrated with a sequencing device. Rather, a stand-alone system or a system integrated with other devices is also possible. A system capable of carrying out a method set forth herein, whether integrated with detection capabilities or not, can include a system controller that is capable of executing a set of instructions to

perform one or more steps of a method, technique or process set forth herein. Optionally, the instructions can further direct the performance of steps for detecting nucleic acids. A useful system controller may include any processor-based or microprocessor-based system, including systems using microcontrollers, reduced instruction set computers (RISC), application specific integrated circuits (ASICs), field programmable gate array (FPGAs), logic circuits, and any other circuit or processor capable of executing functions described herein. A set of instructions for a system controller may be in the form of a software program. As used herein, the terms "software" and "firmware" are interchangeable, and include any computer program stored in memory for execution by a computer, including RAM memory, ROM memory, EPROM memory, EEPROM memory, and non-volatile RAM (NVRAM) memory. The software may be in various forms such as system software or application software. Further, the software may be in the form of a collection of separate programs, or a program module within a larger program or a portion of a program module. The software also may include modular programming in the form of object-oriented programming.

## EXAMPLE 1

Alignment of the locus D18S51

This example describes alignment of the locus D18S51 according to one embodiment. Some loci have flanking sequences which are low-complexity and resemble the STR repeat sequence. This can cause the flanking sequence to be misaligned (sometimes to the STR sequence itself) and thus the allele can be be miscalled. An example of a troublesome locus is D18S51. The repeat motif is **[AGAA]n AAAG AGAGAG**. The flanking sequence is shown below with the low-complexity "problem" sequence underlined:

GAGACCTTGTCTC **(STR)** **GAAAGAAAGAGAAAAAGAAAAGAAA**TAGTAGCAACTGTTAT

If the flanking region immediately adjacent to the STR were used to seed the alignment, k-mers would be generated such as GAAAG, AAAGAA, AGAGAAA, which map to the STR sequence. This deters performance since many possibilities are obtained from the seeding, but most importantly, the approach creates mis-

alignments, such as those shown in Figure 2. In the sequences shown in Figure 2, the true STR sequence is highlighted, the STR sequence resulting from the mis-alignment is underlined and read errors are shown in bold.

For these low-complexity flanks, it was ensured that the seeding regions are not in the low-complexity region by pushing them further away from the STR sequence. While this requires longer reads to call the STR, it ensures high-accuracy and prevents mis-alignment of the flanking region to STR sequence (or other parts of the flank). The low-complexity flank is still aligned to the read to find the ending position of the STR but because the alignment is seeded with high-complexity sequence it has to be in the correct position.

## EXAMPLE 2

Alignment of the locus Penta-D by short STR Sequence Addition

A set of Penta-D sequences tended to have STRs that were 1 nt shorter than expected. Upon further inspection, it was discovered that both flanks contained poly-A stretches and sequencing / amplification errors often removed one of the A's in those stretches. As shown in the sequence below, homopolymeric A stretches are found on both flanks.

```
... CAAGAAAGAAAAAAAAG [AAAGA]n AAAAACGAAGGGGAAAAAAAGAGAAT...
```

A read error causing a deletion in the first flank would yield to two equally viable alignments:

```
read:  ...CAAGAAAGAAAAAAA-GA...
flank: ...CAAGAAAGAAAAAAAAG-      (2 indels)

read:  ...CAAGAAAGAAAAAAAGA...    (2 mismatches)
flank: ...CAAGAAAGAAAAAAAAG
```

Enforcing the base closest to the STR to be a match did not work because one of the flanks in one of the STRs ended up having a SNP in it, causing us to reconsider that method all together. It was discovered that adding just 2 bases of the STR sequence solved the issue:

```
read:  ...CAAGAAAGAAAAAAA-GAA
flank: ...CAAGAAAGAAAAAAAAGAA    (1 indel) ✔
```

```
read:  ...CAAGAAAGAAAAAAAG-AA     (1 indel + 1 mismatch)
flank: ...CAAGAAAGAAAAAAAAGAA
```

## EXAMPLE 3

Analysis of Mixture of DNA Samples

A mixture of samples was analyzed using the methods provided herein to make accurate calls for each locus in a panel of forensic STRs. For each locus, the number reads corresponding to each allele and to each different sequence for that allele were counted.

Typical results are shown in Figure 3. As shown, the bar on the right of each pair represents the actual data obtained, indicating the proportion of reads for each allele. Different shades represent different sequences. Alleles with less than 0.1% of the locus read count and sequences with less than 1% of the allele count are omitted. The bar on the left side of each pair represents the theoretical proportions (no stutter). Different shades represent different control DNA in the input as indicated in the legend. In Figure 3, the x-axis is in order allele, and the Y axis indicates proportion of reads with the indicated allele.

As shown in the Figure, the STR calling approach using the methods presented herein achieved surprisingly accurate calls for each allele in the panel.

## EXAMPLE 4

Analysis of Forensic STR Panel

A panel of 15 different loci were analyzed in 5 different samples. The samples were obtained from Promega Corp, and included samples 9947A, K562, 2800M, NIST: A and B (SRM 2391c). The loci were chosen from the CODIS STR forensic markers and included CSF1PO, D3S1358, D7S820, D16S539, D18S51, FGA, PentaE, TH01, vWA, D5S818, D8S1179, D13S317, D21S11, PentaD and TPOX using the alignment method presented herein. Briefly, the markers were amplified using standard primers, as set forth in Krenke, et al. (2002) *J. Forensic Sci.* 47(4): 773-785. The amplicons were pooled and sequencing data was obtained using 1x460 cycles on a MiSeq sequencing instrument (Illumina, San Diego, CA).

Alignment was performed according to the methods presented herein. As set forth in Fig. 4, 100% concordance for these control samples was shown compared to control data. In addition, this method identified a previously-unknown SNP in one of the samples for marker D8S1179, further demonstrating the powerful tool of sequence-based STR analysis when combined with the alignment methods provided herein.

Throughout this application various publications, patents and/or patent applications have been referenced.

The term comprising is intended herein to be open-ended, including not only the recited elements, but further encompassing any additional elements.

A number of embodiments have been described. Nevertheless, it will be understood that various modifications may be made. Accordingly, other embodiments are within the scope of the following claims.

What is claimed is:

1. A method for determining the length and/or sequence of a polymorphic repetitive DNA element having a repeat region situated between a first conserved flanking region and a second conserved flanking region, the method comprising:

(a) providing a data set comprising at least one sequence read of the polymorphic repetitive DNA element;

(b) providing a reference sequence comprising the first conserved flanking region and the second conserved flanking region;

(c) aligning a portion of the first flanking region of the reference sequence to the sequence read;

(d) aligning a portion of the second flanking region of the reference sequence to the sequence read; and

(e) determining the length and/or sequence of the repeat region;

wherein at least steps (c), (d) and (e) are performed using a suitably programmed computer;

wherein the aligning a portion of the flanking region in one or both of steps (c) and (d) comprises:

(i) determining a location of a conserved flanking region on the read by using exact k-mer matching of a seeding region which overlaps or is adjacent to the repeat region; and

(ii) aligning the flanking region to the sequence read;

wherein the seeding region comprises a high-complexity region of the conserved flanking region, the high-complexity region comprising sequence that is sufficiently distinct from the repeat region so as to avoid mis-alignment.

2. The method of claim 1, further comprising aligning both the flanking sequence and a short adjacent region comprising a portion of the repeat region.

3. The method of claim 1, wherein the high-complexity region comprises a sequence having a diverse mixture of bases.



Date Reçue/Date Received 2020-06-15

4. The method of claim 1, wherein the seeding region avoids low-complexity regions of the conserved flanking region.

5. The method of claim 4, the low-complexity region comprising sequence that substantially resembles that of the repeat sequence.

6. The method of claim 4, the low-complexity region comprising sequence having a mixture of bases with low diversity.

7. The method of claim 1, wherein the seeding region is directly adjacent to the repeat region.

8. The method of claim 1, wherein the seeding region comprises a portion of the repeat region.

9. The method of claim 1, wherein the seeding region is offset from the repeat region.

10. The method of claim 1, wherein the dataset of sequence reads comprises sequence data from a PCR amplicon having a forward and reverse primer sequence.

11. The method of claim 1, wherein the at least one sequence read in the data set comprises a consensus sequence derived from multiple sequence reads.

12. The method of claim 1, wherein providing a reference sequence comprises identifying a locus of interest based upon a primer sequence of a PCR amplicon.

13. The method of claim 1, wherein the at least one sequencing read comprises sequence from a sequencing-by-synthesis (SBS) reaction.



Date Reçue/Date Received 2020-06-15

14. The method of claim 1, wherein the at least one sequencing read comprises sequence from a sequencing-by-ligation reaction.

15. The method of claim 1, wherein the data set is received from a memory.

16. The method of claim 1, wherein the length or sequence of the repeat region is output via a physical or virtual connection, a display or a printer.

17. The method of claim 1, wherein the repeat region is a short tandem repeat (STR).

18. The method of claim 17, wherein the STR is selected from the CODIS autosomal STR loci.

19. The method of claim 17, wherein the STR is selected from the CODIS Y-STR loci.

20. The method of claim 17, wherein the STR is selected from the EU autosomal STR loci.

21. The method of claim 17, wherein the STR is a selected from the EU Y-STR loci.

22. A system for determining the length and/or sequence of a polymorphic repetitive DNA element having a repeat region situated between a first conserved flanking region and a second conserved flanking region, the system comprising:

a processor; and

a program for determining the length and/or sequence of a polymorphic repetitive DNA element, the program comprising instructions for the processor to perform the following steps:



Date Reçue/Date Received 2020-06-15

(a) providing a data set comprising at least one sequence read of the polymorphic repetitive DNA element;

(b) providing a reference sequence comprising the first conserved flanking region and the second conserved flanking region;

(c) aligning a portion of the first flanking region of the reference sequence to the sequence read;

(d) aligning a portion of the second flanking region of the reference sequence to the sequence read; and

(e) determining the length and/or sequence of the repeat region;

wherein the aligning a portion of the flanking region in one or both of steps (c) and (d) comprises:

(i) determining a location of a conserved flanking region on the read by using exact k-mer matching of a seeding region which overlaps or is adjacent to the repeat region; and

(ii) aligning the flanking region to the sequence read;

wherein the seeding region comprises a high-complexity region of the conserved flanking region, the high-complexity region comprising sequence that is sufficiently distinct from the repeat region so as to avoid mis-alignment.

23. The system of claim 22, further comprising aligning both the flanking sequence and a short adjacent region comprising a portion of the repeat region.

24. The system of claim 22, wherein the high-complexity region comprises a sequence having a diverse mixture of bases.

25. The system of claim 22, wherein the seeding region avoids low-complexity regions of the conserved flanking region.

26. The system of claim 25, the low-complexity region comprising sequence that substantially resembles that of the repeat sequence.



Date Reçue/Date Received 2020-06-15

27. The system of claim 25, the low-complexity region comprising sequence having a mixture of bases with low diversity.

28. The system of claim 22, wherein the seeding region is directly adjacent to the repeat region.

29. The system of claim 22, wherein the seeding region comprises a portion of the repeat region.

30. The system of claim 22, wherein the seeding region is offset from the repeat region.

31. The system of claim 22, wherein the dataset of sequence reads comprises sequence data from a PCR amplicon having a forward and reverse primer sequence.

32. The system of claim 22, wherein the at least one sequence read in the data set comprises a consensus sequence derived from multiple sequence reads.

33. The system of claim 22, wherein providing a reference sequence comprises identifying a locus of interest based upon the primer sequence of the PCR amplicon.

34. The system of claim 22, wherein the at least one sequencing read comprises sequence from a sequencing-by-synthesis (SBS) reaction.

35. The system of claim 22, wherein the at least one sequencing read comprises sequence from a sequencing-by-ligation reaction.

36. The system of claim 22, wherein the data set is received from a memory.



Date Reçue/Date Received 2020-06-15

37. The system of claim 22, wherein the length or sequence of the repeat region is output via a physical or virtual connection, a display or a printer.

38. The system of claim 22, wherein the repeat region is a short tandem repeat (STR).

39. The system of claim 38, wherein the STR is selected from the CODIS autosomal STR loci.

40. The system of claim 38, wherein the STR is selected from the CODIS Y-STR loci.

41. The system of claim 38, wherein the STR is selected from the EU autosomal STR loci.

42. The system of claim 38, wherein the STR is a selected from the EU Y-STR loci.

43. A computer-implemented method for determining the length and/or sequence of a polymorphic repetitive DNA element having a repeat region situated between a first conserved flanking region and a second conserved flanking region, the method comprising:

(a) providing a data set comprising a plurality of sequence reads, wherein the plurality of sequence reads is next-generation sequencing (NGS) whole genome sequence data;

(b) providing a reference sequence comprising the first conserved flanking region and the second conserved flanking region;

(c) comparing a portion of the first conserved flanking region of the reference sequence to the plurality of sequence reads to align one or more sequence reads to the portion of the first conserved flanking region;

(d) comparing a portion of the second conserved flanking region of the reference sequence to the plurality of sequence reads to align the one or more sequence reads to the portion of the second conserved flanking region; and



Date Reçue/Date Received 2020-06-15

(e) determining the length and/or sequence of the repeat region;

wherein at least steps (c), (d) and (e) are performed using a suitably programmed computer, and wherein the comparing the portion of the first or second conserved flanking region of the reference sequence comprises:

(i) performing, on the plurality of sequence reads, exact k-mer matching of a seeding region which overlaps or is adjacent to the repeat region to determine a location of a conserved flanking region on the one or more sequence reads; and

(ii) aligning the portion of the first or second conserved flanking region of the reference sequence to the one or more sequence reads.

44. The method of claim 43, further comprising aligning both the flanking region and a short adjacent region comprising a portion of the repeat region.

45. The method of claim 43, wherein the seeding region comprises a high-complexity region of the conserved flanking region.

46. The method of claim 45, the high-complexity region comprising a sequence that is sufficiently distinct from the repeat region so as to avoid mis-alignment.

47. The method of claim 45, wherein the high-complexity region comprises a sequence having a diverse mixture of bases.

48. The method of claim 43, wherein the seeding region avoids low-complexity regions of the conserved flanking region.

49. The method of claim 48, the low-complexity regions comprising a sequence that substantially resembles that of the repeat region.

50. The method of claim48, the low-complexity regions comprising a sequence having a mixture of bases with low diversity.



Date Reçue/Date Received 2020-06-15

51. The method of claim 43, wherein the seeding region is directly adjacent to the repeat region.

52. The method of claim 43, wherein the seeding region comprises a portion of the repeat region.

53. The method of claim 43, wherein the seeding region is offset from the repeat region.

54. The method of claim 43, wherein the data set comprises sequence data from a PCR amplicon having a forward and reverse primer sequence.

55. The method of claim 43, wherein the plurality of sequence reads comprises a consensus sequence derived from multiple sequence reads.

56. The method of claim 54, wherein providing a reference sequence comprises identifying a locus of interest based upon the primer sequence of the PCR amplicon.

57. The method of claim 43, wherein the plurality of sequence reads comprises sequence reads from a sequencing-by-synthesis (SBS) reaction.

58. The method of claim 43, wherein the plurality of sequence reads comprises sequence reads from a sequencing-by-ligation reaction.

59. The method of claim 43, wherein the data set is received from a memory.

60. The method of claim 43, wherein the length or sequence of the repeat region is output via a physical or virtual connection, a display or a printer.

61. The method of claim 43, wherein the repeat region is a short tandem repeat (STR).



Date Reçue/Date Received 2020-06-15

62. The method of claim 61, wherein the STR is selected from the CODIS autosomal STR loci.

63. The method of claim 61, wherein the STR is selected from the CODIS Y-STR loci.

64. The method of claim 61, wherein the STR is selected from the EU autosomal STR loci.

65. The method of claim 61, wherein the STR is a selected from the EU Y-STR loci.

66. A system for determining the length and/or sequence of a polymorphic repetitive DNA element having a repeat region situated between a first conserved flanking region and a second conserved flanking region, the system comprising:

a processor; and

a program for determining the length of a polymorphic repetitive DNA element, the program comprising instructions for:

(a) providing a data set comprising a plurality of sequence reads, wherein the plurality of sequence reads is NGS whole genome sequence data;

(b) providing a reference sequence comprising the first conserved flanking region and the second conserved flanking region;

(c) comparing a portion of the first conserved flanking region of the reference sequence to the plurality of sequence reads to align one or more sequence reads to the portion of the first conserved flanking region;

(d) comparing a portion of the second conserved flanking region of the reference sequence to the plurality of sequence reads to align the one or more sequence reads to the portion of the second conserved flanking region; and

(e) determining the length and/or sequence of the repeat region,

wherein the comparing the portion of the first or second conserved flanking region of the reference sequence comprises:



Date Reçue/Date Received 2020-06-15

(i) performing, on the plurality of sequence reads, exact k-mer matching of a seeding region which overlaps or is adjacent to the repeat region to determine a location of a conserved flanking region on the one or more sequence reads, and

(ii) aligning the portion of the first or second conserved flanking region of the reference sequence to the one or more sequence reads.



Date Reçue/Date Received 2020-06-15

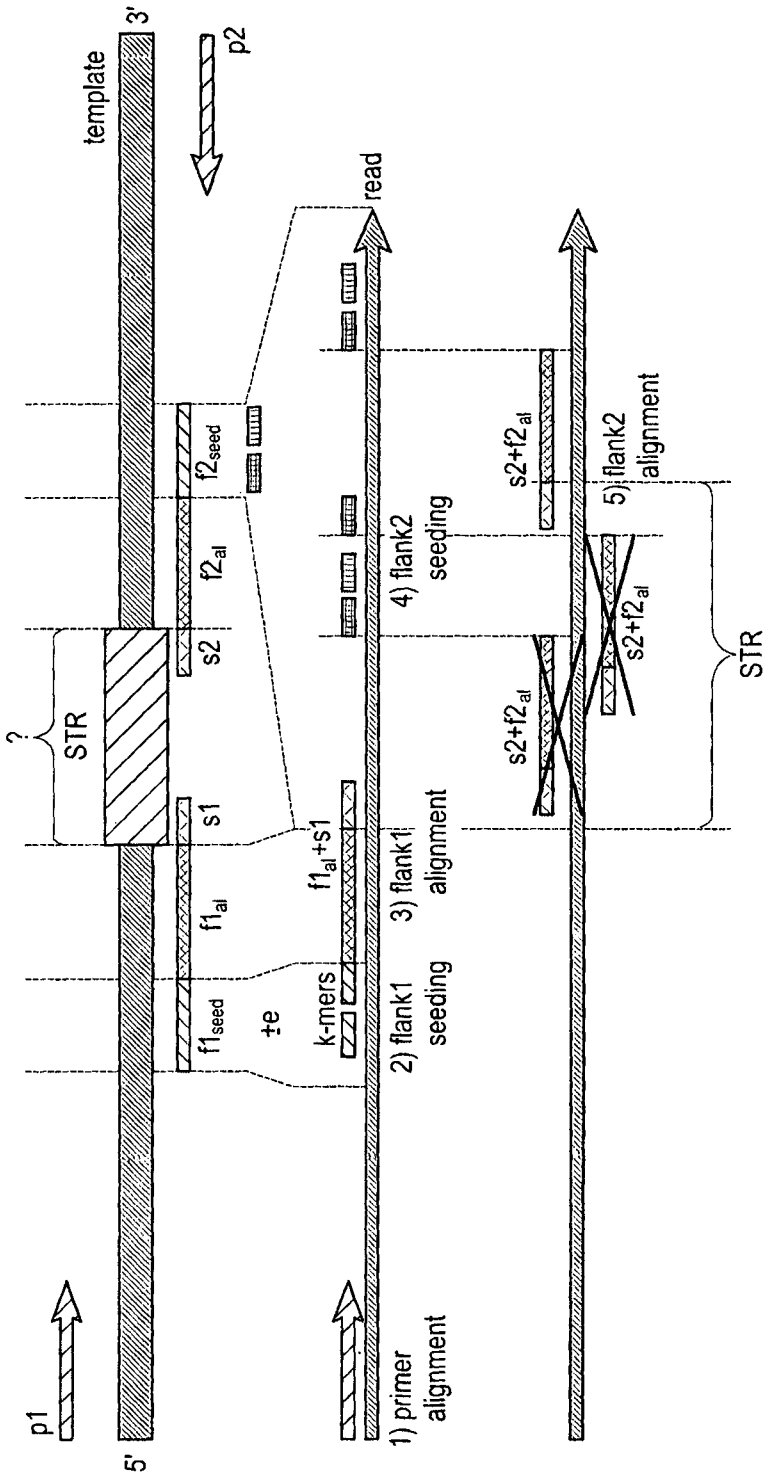

template
3'
p2
read
?
STR
s2
$f2_{al}$
$f2_{seed}$
s1
$f1_{al}$
$f1_{seed}$
±e
k-mers
$f1_{al}$+s1
$s2+f2_{al}$
$s2+f2_{al}$
$s2+f2_{al}$
STR
p1
5'
1) primer alignment
2) flank1 seeding
3) flank1 alignment
4) flank2 seeding
5) flank2 alignment

Fig. 1

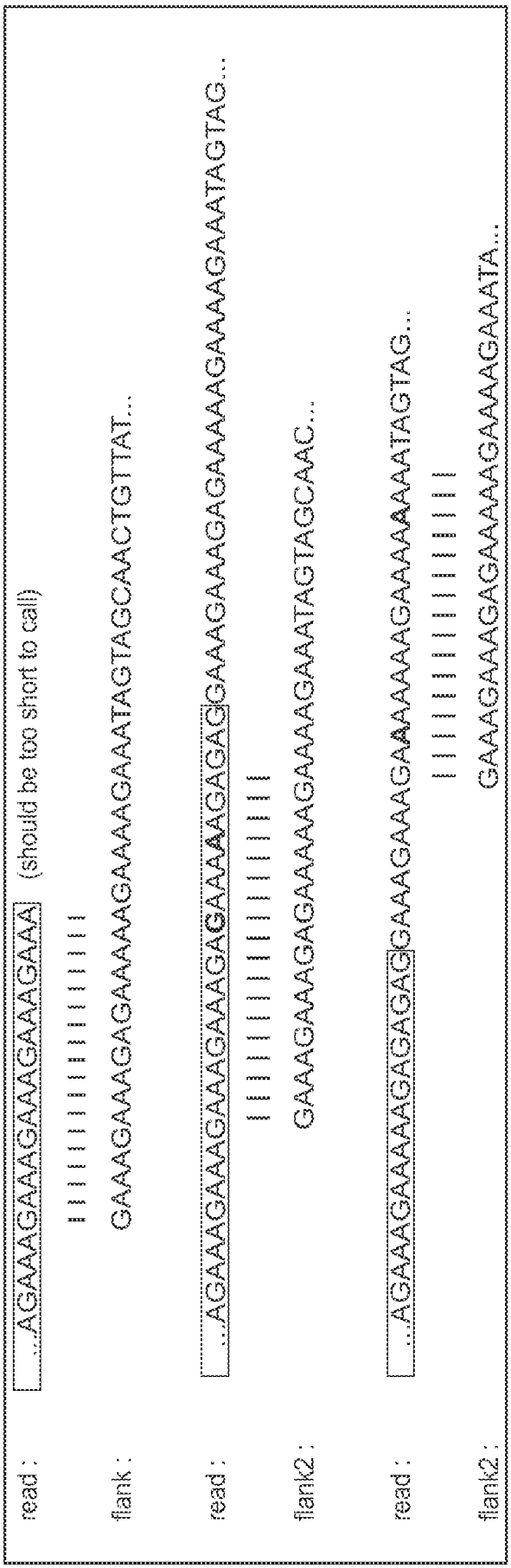

read :
...AGAAAGAAAGAAAGAAA (should be too short to call)
flank :
GAAAGAAAGAGAAAAAGAAATAGTAGCAACTGTTAT...
read :
...AGAAAGAAAGAAAGAGAAAAAGAGAGGAAAGAGAAAAGAAAAGAAATAGTAG...
flank2 :
GAAAGAAAGAGAAAAAGAAAAGAAATAGTAGCAAC...
read :
...AGAAAGAAAAAGAGAGAGGAAAGAAAGAAAAAAAAAAATAGTAG...
flank2 :
GAAAGAAAGAGAAAAAGAAAAGAAATA...

Fig. 2

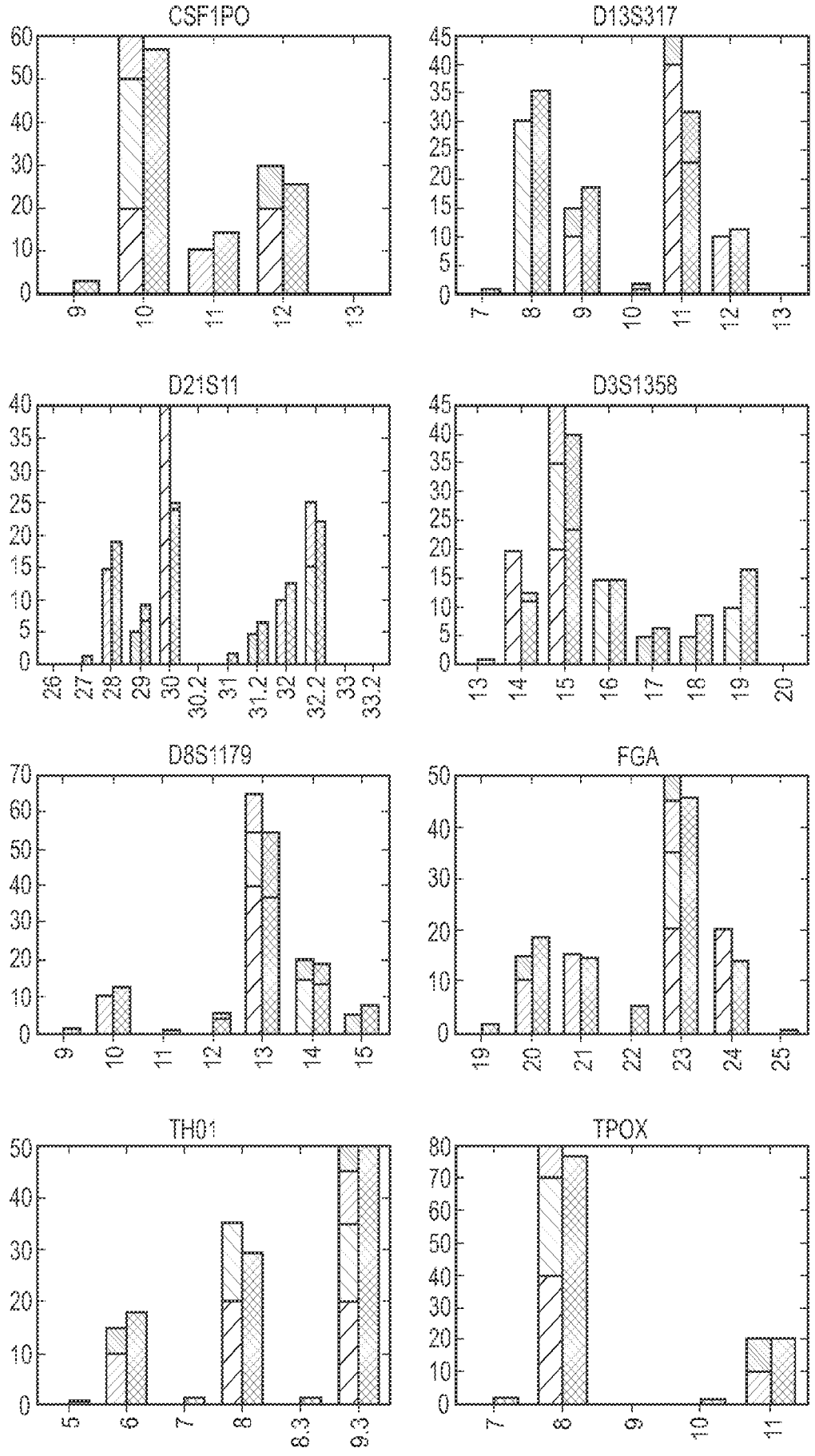



Fig. 3

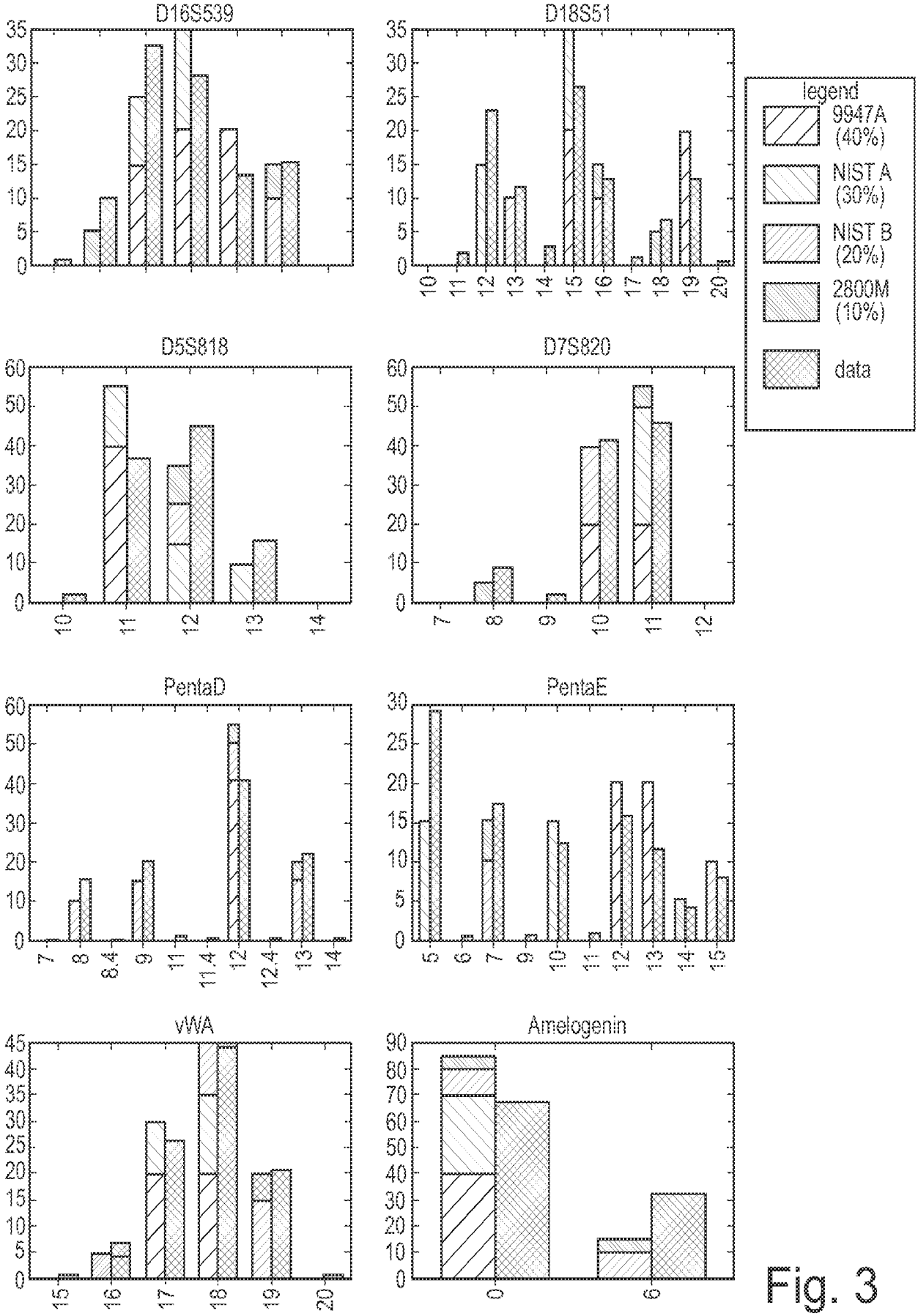
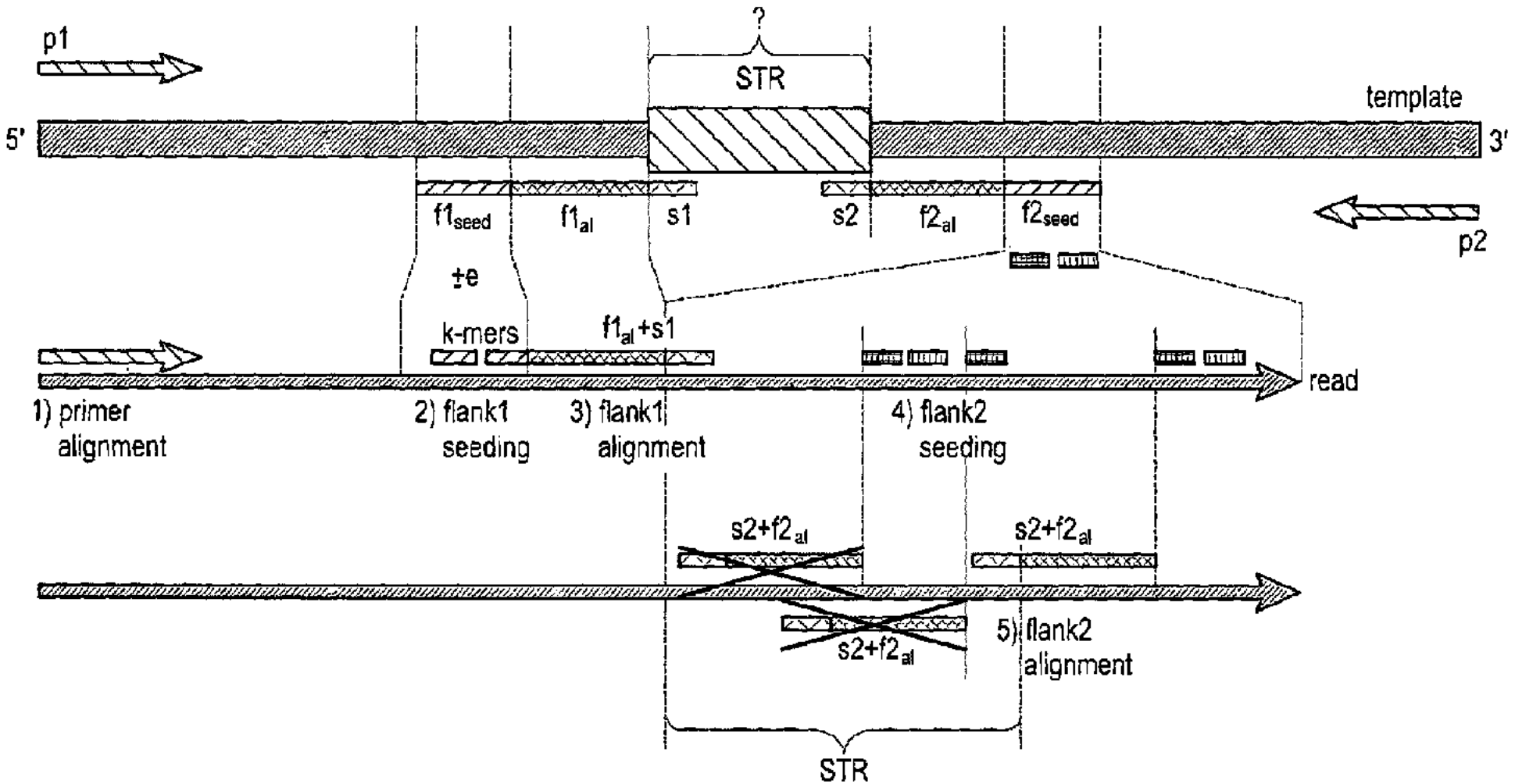

D16S539
D18S51
D5S818
D7S820
PentaD
PentaE
vWA
Amelogenin
legend
9947A (40%)
NIST A (30%)
NIST B (20%)
2800M (10%)
data

Fig. 3

| DNA Sample | CSF1PO | D3S1358 | D7S820 | D16S539 | D18S51 | FGA | PentaE | TH01 | vWA | D5S818 | D8S1179 | D13S317 | D21S11 | PentaD | TPOX |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9947A | 10, 12 | 14, 15 | 10, 11 | 11, 12 | 15, 19 | 23, 24 | 12, 13 | 8, 9.3 | 19, 20 | 11 | 13, 13[1] | 11 | 30 | 12 | 8 |
| 2800M | 12 | 17, 18 | 8, 11 | 9, 13 | 16, 18 | 20, 23 | 7, 14 | 6, 9.3 | 16, 19 | 12 | 14, 15 | 9, 11 | 29, 31.2 | 12, 13 | 11 |
| NIST A | 10 | 15, 16 | 11 | 10, 11 | 12, 15 | 21, 23 | 5, 10 | 8, 9.3 | 18, 19 | 11, 12 | 13, 14 | 8 | 28, 32.2 | 9, 13 | 8 |
| NIST B | 10, 11 | 15, 19 | 10 | 10, 13 | 13, 16 | 20, 23 | 7, 15 | 6, 9.3 | 17, 18 | 12, 13 | 10, 13 | 9, 12 | 32, 32.2 | 8, 12 | 8, 11 |
| NIST C | 10, 12 | 16, 18 | 10, 12 | 10 | 16, 19 | 24, 26 | 12, 13 | 6, 8 | 16, 18 | 10, 11 | 10, 17 | 11 | 29, 30 | 10, 11 | 11 |
| | | | | | | | | | | | | | | | |
| [1]Discovered SNP in half of the repeats: 46% [TCTA]13,56% [TCTA][TCTG][TCTA]11 | | | | | | | | | | | | | | | |

Fig. 4


p1
?
STR
template
5'
3'
f1seed
f1al
s1
s2
f2al
f2seed
p2
±e
k-mers
f1al+s1
read
1) primer alignment
2) flank1 seeding
3) flank1 alignment
4) flank2 seeding
s2+f2al
s2+f2al
s2+f2al
5) flank2 alignment
STR