



US 20110263437A1

(19) **United States**(12) **Patent Application Publication**
FANG et al.(10) **Pub. No.: US 2011/0263437 A1**(43) **Pub. Date: Oct. 27, 2011**(54) **ANALYSIS OF Y-CHROMOSOME STR
MARKERS***G01N 27/447* (2006.01)*C12Q 1/68* (2006.01)*C07H 21/04* (2006.01)(75) Inventors: **Rixun FANG**, Menlo Park, CA
(US); **Manohar Furtado**, San
Ramon, CA (US); **Manfred**
Kayser, Rotterdam (NL); **Kaye**
Ballantyne, Hawthorn East (AU)(52) **U.S. Cl. 506/7**; 435/6.11; 536/23.1; 536/24.33;
250/282; 204/451; 204/456(73) Assignee: **LIFE TECHNOLOGIES
CORPORATION**, Carlsbad, CA
(US)(21) Appl. No.: **12/880,040**(22) Filed: **Sep. 10, 2010****Related U.S. Application Data**(60) Provisional application No. 61/241,778, filed on Sep.
11, 2009, provisional application No. 61/367,346,
filed on Jul. 23, 2010, provisional application No.
61/379,340, filed on Sep. 1, 2010.**Publication Classification**(51) **Int. Cl.**
C40B 30/00 (2006.01)
G01N 33/559 (2006.01)
H01J 49/26 (2006.01)(57) **ABSTRACT**

The methods and compositions provided herein relate to the discovery of 13 STR are markers found on the human Y chromosome with surprisingly high mutation rates when compared with 173 other Y-STR markers known today, including those that are in common use in forensics. In addition to theoretical expectations based on elevated mutation rates, these 13 rapidly-mutating (RM) Y-STRs proved to be suitable for differentiating between closely related males, as well as between more distant male relatives and much more so than the most-commonly used Y-STRs in forensics. Our new set of RM-Y-STRs is expected to overcome the current dilemma of Y-chromosome analysis in forensic applications from current male lineage identification towards male individual identification in many cases, due to their extraordinary mutation properties as discovered here. Embodiments of the invention include methods for allelic determination of rapidly-mutating Y-STR markers, amplification primers for the analysis of rapidly-mutating Y-STR markers, allelic ladders for analysis of rapidly-mutating Y-STR markers, and kits for the analysis of rapidly-mutating Y-STR markers.

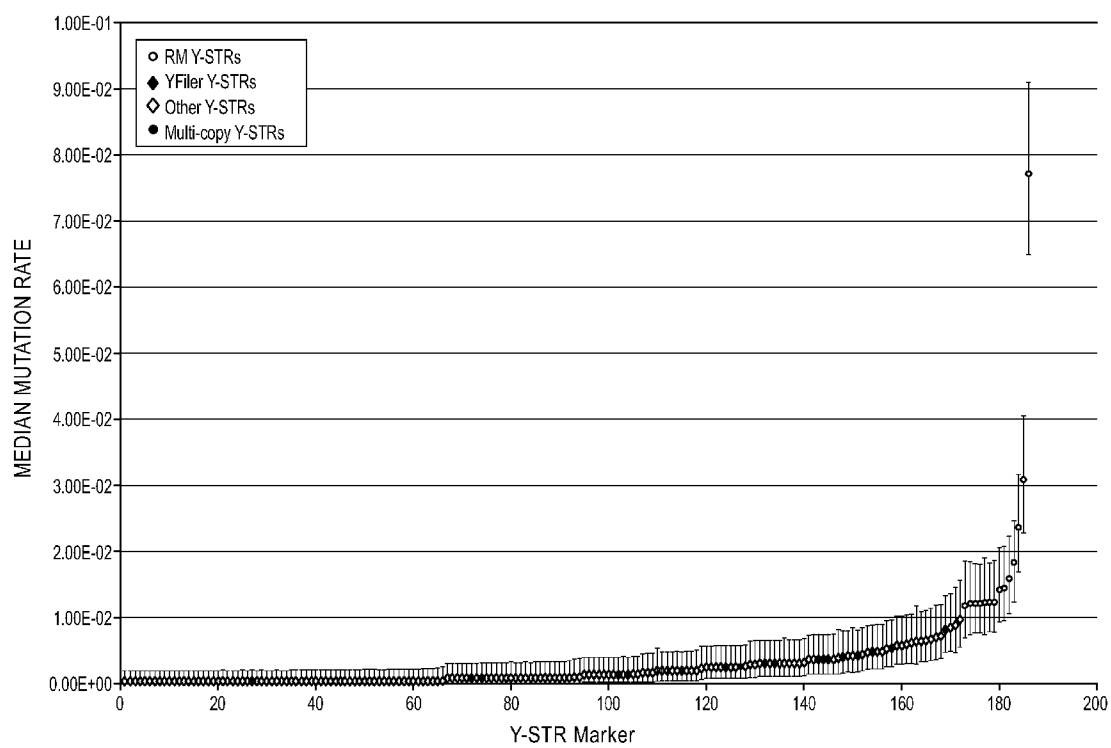


FIG. 1

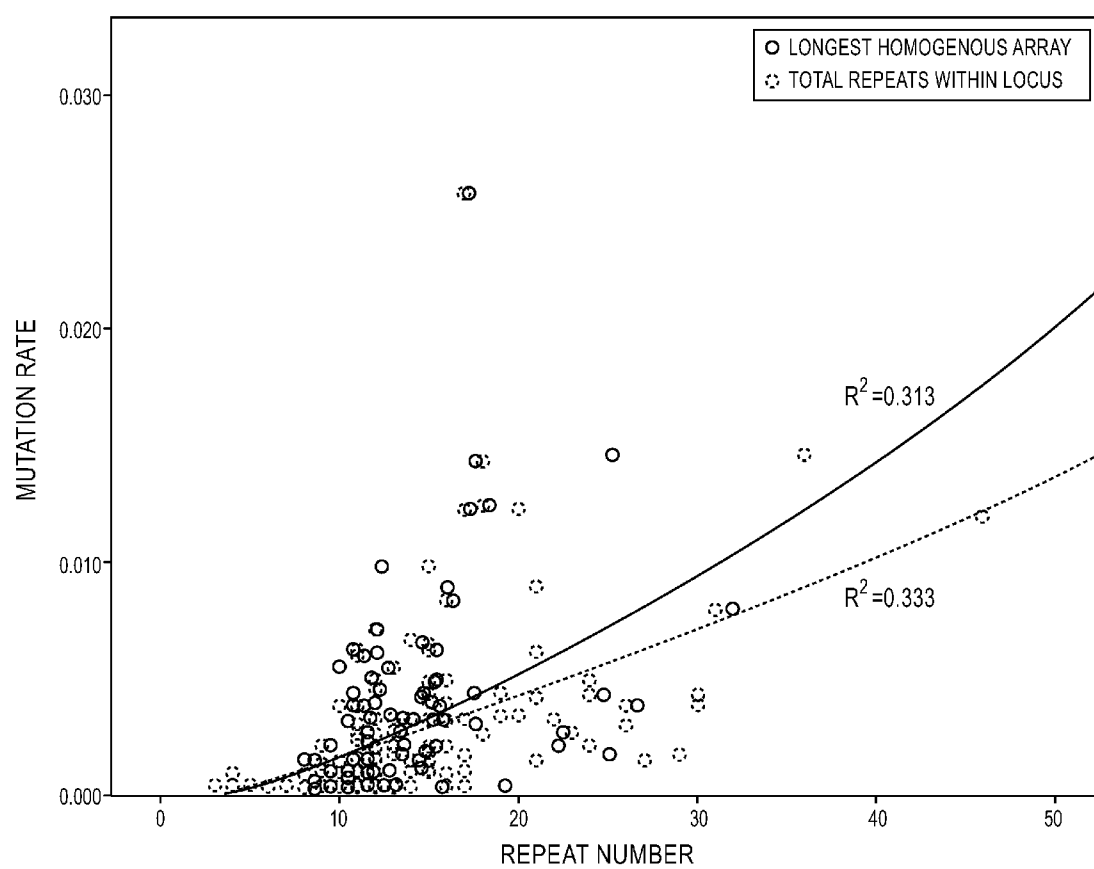


FIG. 2

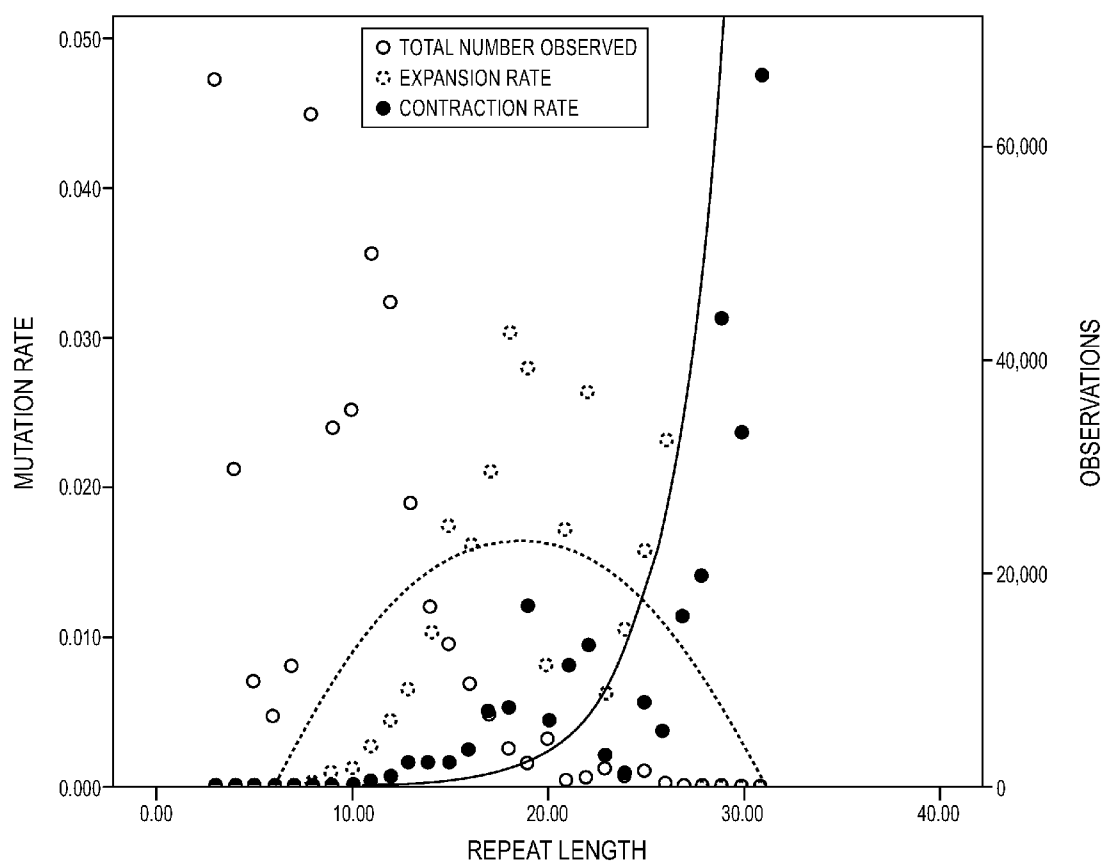


FIG. 3

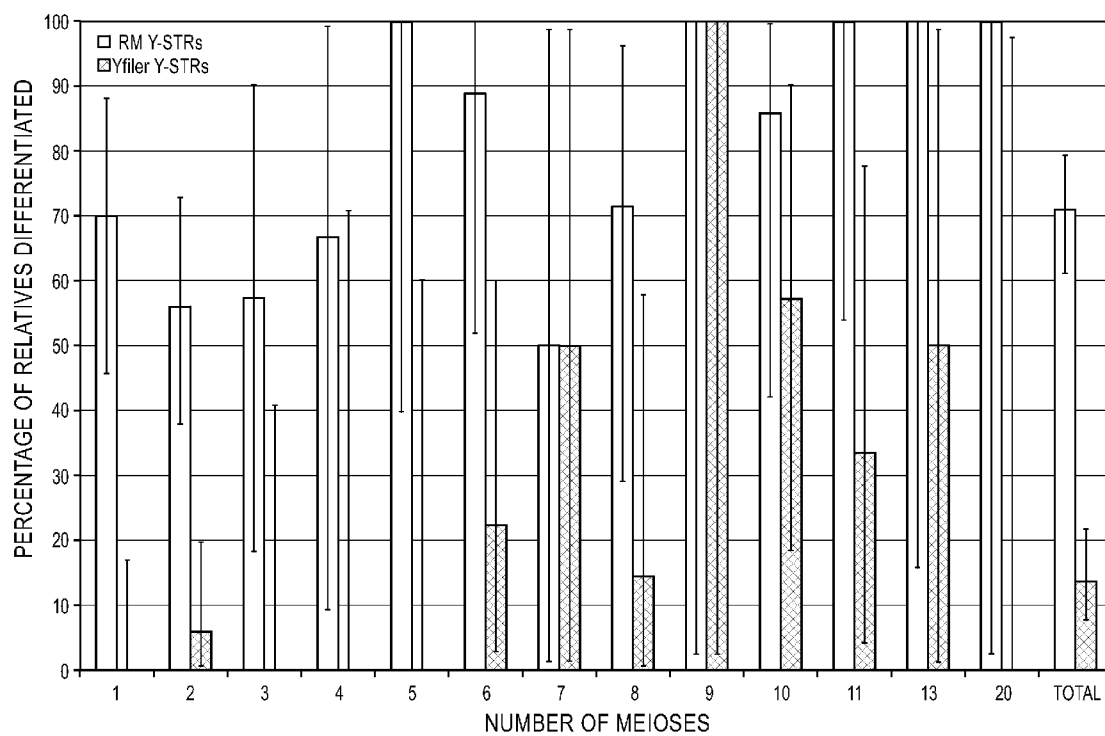


FIG. 4

ANALYSIS OF Y-CHROMOSOME STR MARKERS

[0001] This application claims priority to U.S. Provisional Application No. 61/241,778, filed Sep. 11, 2009, U.S. Provisional Application No. 61/367,346 filed Jul. 23, 2010, and to U.S. Provisional Application No. 61/379,340 filed Sep. 1, 2010. Application Nos., Ser. Nos. 61/241,778, 61/367,346 and 61/379,340 are incorporated by reference herein in their entirety for any purpose.

FIELD

[0002] Embodiments of the subject inventions are in the field of the forensic analysis of DNA.

BACKGROUND

[0003] The use of STR markers has become a standard tool in the analysis of DNA found at crime scenes. In most cases, the use of autosomal STR markers are used because, in part, of the high level of polymorphisms within most populations. For example, the 13 CODIS loci that are the standard for databasing criminal suspect in DNA in the United States are autosomal STR markers. In many cases with mixed stains from male and female contributors, particularly rape cases, forensic investigators must analyze genetic markers found on the Y chromosome to identify the male component usually belonging to the perpetrator of the crime. This is because in such cases, the autosomal STR markers are not informative due to profile overlap between e.g. female victim DNA and male perpetrator DNA. Although there are technical possibilities (i.e. differential lysis) to preferentially access male DNA, such techniques are often not successful. Because female DNA lacks a Y chromosome, the analysis of Y chromosomal markers can be used in samples that contained high levels of female DNA relative to the male DNA in the sample. Analyzing the Y chromosomal DNA hence excludes the complicating artifacts caused by the excess female source DNA.

[0004] The non-recombining nature allows the use of Y chromosome markers for male lineage identification, i.e. groups of males that are paternally related and hence share the same Y-STR haplotype i.e. based on currently-used Y-STR markers in forensics. Male lineage identification has become a valuable tool in forensic genetics to exclude males. However, in cases of non-exclusion (i.e. matching Y-STR profiles) no individual-based statement can be made based on the currently-available Y-STR markers because the same probability of having donated the crime scene sample applies to a male suspect and all his male relatives. This clearly is a limitation in forensic application where individual-based conclusions are anticipated. However, mutation events can occur at Y-STR markers. These mutations in the Y-STR marker can in principle enable the investigator to distinguish between closely related male relatives, and also between more distantly related males, provided such mutations occur in high-enough frequencies to be observable in a give pair of male relatives. Mutations in the currently available Y-STR markers are fairly infrequent events, occurring on the order of about 0.1 to 0.4% (1-4 changes per thousand generational events per each Y-STR locus). Thus even when relatively large numbers of Y-STR markers, i.e. those 17 markers applied to forensic applications today, are used the probability of distinguishing between male relatives is still remote.

However, if enough Y-STRs markers that mutate more rapidly than the currently-known Y-STRs would be available, it can be expected that closely related males as well as distantly related males become differentiable based on Y-STR mutations towards male individual identification as anticipated in forensic applications.

[0005] The inventors have discovered a subset of thirteen Y-STR markers that have a significantly higher mutation rate than most Y-STR markers including those that are in general use. This finding is expected to revolutionize Y chromosomal applications in forensic biology, from previous male lineage differentiation methods. This finding also leads the way for male individual identification. Thus, by using one or more, by using two or more of such rapidly-mutating Y-STR markers (RM Y-STRs), the ability to distinguish between close and distantly related male relatives is significantly increased.

SUMMARY

[0006] Certain embodiments of the invention include methods of identifying an individual by determining the allele of at least 2 Y-STR markers selected from the group consisting of the rapidly-mutating Y-STR markers: DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627. In some embodiments of the subject methods, the alleles can be identified by PCR. In some embodiments of the subject methods, the alleles can be identified by mass spectroscopy. The PCR can be multiplexed PCR so as to co-amplify the at least 2 of the rapidly-mutating Y-STR markers. Certain embodiments of the invention include set of amplification primer pairs comprising primers for the amplification of at least 2 Y-STR markers selected from the group consisting of DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627. The primers set can co-amplify at least 2-13 of the rapidly-mutating Y-STR markers. In certain embodiments the primer set can co-amplify autosomal STR markers in addition to rapidly-mutating Y-STR markers. In some embodiments, the autosomal STRs can be selected from the group consisting of D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820, D16S539, TH01, TPOX, and CSF1 PO. In some embodiments the primers can be labeled with a fluorescent dye. Other embodiments provided are allelic ladder size standard for calling one or more alleles of an STR from at least 2 of the Y-STR markers selected from the group consisting of DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627. Other embodiments provided are kits for identifying the allele of at least 2 Y chromosome STRS markers, wherein the markers are selected from the group consisting of DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627, the kit comprising primers for the amplification of at least 2 rapidly-mutating Y-STR markers, and an allelic ladder representative of the selected markers.

BRIEF DESCRIPTION OF THE DRAWINGS AND TABLES

[0007] FIG. 1. Mutation rates of 186 Y-STR markers established from father-son pair analysis. Distribution of 186 Y-STR markers according to their Bayesian-based mutation

rates (with credible intervals) estimated from analyzing up to 1966 DNA confirmed father-son pairs per each marker. The 13 rapidly-mutation (RM) Y-STR markers ascertained for further family/pedigree analysis are highlighted in red, and the commonly-used 17 Yfiler Y-STRs are in green. Multi-copy Y-STRs are noted with a black insert diamond.

[0008] FIG. 2. Correlation between the length of the longest homogeneous array, or the total number of repeats within a locus, and the allele-specific mutation rate from 267 Y-STR loci. Although the number of repeats present within a locus “longest homogenous array can be used to predict mutability, the total number of all repeats present within the locus has higher predictive value.

[0009] FIG. 3. Relationship between total number of repeats and mutation direction and rate from 267 Y-STR loci. Repeat loss mutations (contractions) displayed an exponential relationship with the total number of repeats, with increasing rates of loss rates at loci with higher numbers of repeats. Repeat gain mutations (expansions) showed a weak quadratic function, with a peak in gain rate at 20 total repeats.

[0010] FIG. 4. Male relative differentiation with newly-identified 13 RM Y-STRs and commonly-used 17 Yfiler Y-STRs. Results from differentiating between male relatives from analyzing 103 pairs from 80 male pedigrees, sorted according to the number of generations separating pedigree members, based on 13 RM Y-STRs (in red) and 17 Yfiler Y-STRs (in blue). Error bars represent 95% binomial confidence intervals. Note that these samples are independent from the father-son pairs initially used to establish the Y-STR mutation rates.

[0011] Table 1. Mutation rate estimations from the posterior distributions (medians and 95% credible intervals) of 186 Y-STR markers from analyzing up to 1966 DNA-confirmed father-son pairs. Markers with median mutation rates above 10^{-2} (the RM Y-STR set) are highlighted. Additionally included are marker repeat structures, number of gains/losses, total mutations and total number of father-son transmissions observed. PCR primers, PCR annealing temperature and locus assignment to the 54 multiplexes and three RM Y-STR multiplexes used for genotyping are included.

[0012] Table 2. Details of the 924 mutations observed among 120 Y-STR markers from screening a total of 352,999 meiotic transfers at 186 Y-STR markers. The repeat structure of both the father and son’s alleles at the mutated Y-STR are given where possible. In the case of multi-copy markers with multiple variable segments within the amplicon, total repeat numbers or amplicon size is given in the absence of sequence information. The age of the father at the time of the son’s birth is given, as is an individual pair reference.

[0013] Table 3. Comparison of 13 rapidly mutating RM Y-STRs and 17 Yfiler Y-STRs to differentiate between male relatives by one or more mutations from analyzing 103 pairs from 80 male pedigrees according to the number of generations separating members of the same pedigree.

DEFINITIONS

[0014] A “mutation” in a Y-STR marker is a change in the length of the repeat region of an STR marker or a change in the length (i.e., number) of the bases that are interspersed with the repeat units. For example, the addition of one more repeat unit is mutation resulting in the appearance of a new allele. In another example, the addition of a single base within a single repeat unit is also a mutation resulting in the appearance of a new allele. Such changes can result from the addition or

deletion of one or more repeat units (or fractions thereof). Such sequence changes are readily detected by methods of analysis that are capable of detecting variations in nucleic acid sequence length or nucleic acid base order.

[0015] The term “rapidly-mutating Y-STR marker” (RM Y-STRs) as used herein refers to the following 11 Y-STR markers: DYF387S1, DYF399S1, DYF404S1, DYS449, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627.

[0016] As used herein, the term “allelic ladder” refers to a standard size marker consisting of amplified alleles from a given STR locus or a size standards equivalent in size (or electrophoretic mobility) to the amplified alleles from a given STR locus. An allelic ladder can comprise a size standard for one or more alleles of a given STR marker. An allelic ladder can include alleles from different STR markers. The size standards in an allelic ladder can be labeled with a detectable label, e.g., a fluorescent dye.

[0017] The term “Y-STR marker” as used herein refers to an STR marker that is present on the non-recombining part of the human Y chromosome. Over 250 such Y-STR markers exist based on current knowledge. Y-STR markers are well-known to the person ordinary skill in the art. Database of Y-STR marker are publicly available, for example, at web sites, www.usystrdatabase.org and www.yhrd.org

[0018] The term “STR” as used herein refers to regions of genomic DNA which contain short, repetitive sequence elements. The sequence elements that are repeated are not limited to but are generally three to seven base pairs in length. Each sequence element is repeated at least once within an STR and is referred to herein as a “repeat unit.” The term STR also encompasses a region of genomic DNA wherein more than a single repeat unit is repeated in tandem or with intervening bases, provided that at least one of the sequences is repeated at least two times in tandem.

[0019] The term “Primer” as used herein refers to a single-stranded oligonucleotide or DNA fragment that hybridizes with a DNA strand of a locus in such a manner that the 3' terminus of the primer can act as a site of polymerization and extension using a DNA polymerase enzyme. Primers can also DNA analogs in additions to or instead of naturally occurring DNA, e.g., LNAs, base analogs, and the like. “Primer pair” refers to two primers comprising a primer 1 that hybridizes to a single strand at one end of the DNA sequence to be amplified, and a primer 2 that hybridizes with the other end on the complementary strand of the DNA sequence to be amplified. “Primer site” refers to the area of the target DNA to which a primer hybridizes

[0020] As used herein, the terms “a,” “an,” and “the” and similar referents used herein are to be construed to cover both the singular and the plural unless their usage in context indicates otherwise. Accordingly, the use of the word “a” or “an” when used in conjunction with the term “comprising” in the claims or specification may mean “one,” but it is also consistent with the meaning of “one or more,” “at least one,” and “one or more than one.” It is further noted that the claims may be drafted to exclude any optional element. As such, this statement is intended to serve as antecedent basis for use of such exclusive terminology as “solely,” “only” and the like in connection with the recitation of claim elements, or use of a “negative” limitation.

[0021] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as is commonly understood by one of ordinary skill in the art to which these

inventions belong. All patents, patent applications, published applications, treatises and other publications referred to herein, both supra and infra, are incorporated by reference in their entirety. If a definition and/or description is set forth herein that is contrary to or otherwise inconsistent with any definition set forth in the patents, patent applications, published applications, and other publications that are herein incorporated by reference, the definition and/or description set forth herein prevails over the definition that is incorporated by reference. The citation of any publication is for its disclosure prior to the filing date and should not be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention.

Description of Certain Specific Embodiments

[0022] Applicants have identified mutation rates for numerous Y-STRs by examining three areas: i) the lack of knowledge on Y-STR mutability based on a reasonably large number of loci as required for evolutionary and genealogical applications, ii) the limited knowledge on the molecular basis of Y-STR mutability, and iii) the lack of Y-STRs for familial differentiation in forensic, genealogical, and particular population applications.

[0023] In ~2000 DNA-confirmed father-son pairs. Table 1 presents the mutation rates and characteristics for 186 Y-STR markers. Included are mutation rate estimates, most determined for the first time. Also evaluated were the diversity and DNA sequence data generated for all loci to investigate the underlying causes of Y-STR mutability. The suitability of the identified most mutable Y-STRs for male relative differentiation and their implication for Y-chromosome applications in forensic science have been tested and resulted in the identification of 13 rapidly mutating Y-STR (RM-Y-STR) markers.

[0024] The 13 Y-STR markers were found to have a mutational rate that is substantially higher than the 173 other Y-STRs tested. These rapidly-mutating markers are DYS387S1, DYS399S1, DYS403S1, DYS404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627. The mutation rates for these 13 RM-Y-STRs are all well above 10^{-2} , whereas all other 173 Y-STRs (94% of the loci tested) have mutation rates well below 10^{-2} (usually 10^{-3} and lower) (FIG. 1). In particular, the locus-specific mutation rates of the 13 RM Y-STRs range from 0.0116 to 0.0744. In comparison, the 17 Y-STRs included in the AmpF/STR® YFiler™ PCR Amplification kit (YFiler Kit, sold by Applied Biosystems/Life Technologies, Foster City, Calif. USA, namely DYS456, DYS389I, DYS390, DYS389II, DYS458, DYS19, DYS385 a/b*, DYS393, DYS391, DYS439, DYS635, DYS392, Y GATA H4, DYS437, DYS438, DYS448) have locus-specific mutation rates ranging from 0.0002 to 0.0065 as established recently based on a large number of >135,000 meiotic transfers (Goedbloed et al. 2009). Hence, Applicants have surprisingly discovered that the 13 RM-Y-STRs mutate 60-11 times more rapidly than YFiler kit Y-STRs that are most commonly used in forensic applications today. The surprisingly high mutation rate in these RM-Y-STR markers permits the increased likelihood of distinguishing between male members of the same paternal genetic lineage. The likelihood of discrimination between members of the same male lineage is even greater when multiple rapidly-mutating Y-STR markers are employed. Various embodiments of the invention provided herein include methods, reagents, and kits for determining the specific allele of one or more, of two or more, of three or

more, of four or more, of five or more, and so on, of the subject rapidly-mutating Y-STR markers in a given sample for analysis.

[0025] Provided herein are various methods for determining the specific allele of one or more of the rapidly-mutating Y-STR markers. The specific alleles of the rapidly-mutating Y-STR markers can be determined using essentially the same methods and technologies that are used for the determination of alleles other types of STR markers. Such methods and technologies can readily be adapted by the person skilled in the art so as to be suitable for use in the allele determination of the rapidly-mutating Y-STR markers. Examples of such technology include DNA sequencing and sequence specific amplification techniques such as PCR, used in conjunction with detection technologies such as electrophoresis, mass spectroscopy, and the like. In some embodiments, PCR amplification products may be detected by fluorescent dyes conjugated to the PCR amplification primers, for example as described in PCT patent application WO 2009/059049. PCR amplification products can also be detected by other techniques, including, but not limited to, the staining of amplification products, e.g. silver staining and the like.

[0026] The specific allele of a given rapidly-mutating Y-STR marker can also be determined by any of a variety of DNA sequencing techniques that are widely available, e.g., Sanger sequencing, pyrosequencing, Maxim and Gilbert sequencing, and the like. Numerous automated DNA sequencing techniques are commercially available, the applied Biosystems 3130, the applied Biosystems 3100, the Illumina Genome Analyzer, the Applied Biosystems SOLiD system, the Roche Genome Sequencer Flx system and the like.

[0027] DNA for analysis using the subject methods and compositions can be obtained from a variety of sources. DNA can be obtained at crime scenes, e.g., semen recovered from a rape victim. Additionally, DNA for analysis can be obtained directly from male subjects for the purpose of generating a database of allelic information (for subsequent analysis) or can be obtained from identified suspects.

[0028] DNA for analysis can be quantified prior to allelic analysis, thereby providing for more accurate allele calling. DNA quantity in a sample may be determined by many techniques known to the person skilled in the art, e.g., real time PCR. It is of interest to quantify the Y chromosomal DNA present in a sample for analysis prior to performing allelic analysis for Y-chromosomal STR markers, including rapidly-mutating Y-chromosomal STR markers. Autosomal DNA in the sample may also be quantitated, thereby providing a method for determining the background amount of female DNA present in a mixed sample, such as those samples recovered in rape cases.

[0029] A Y chromosomal haplotype can be established by determining the specific alleles present on a plurality of Y-STR markers. In general, the more rapidly a Y-STR marker mutates, the greater the probability of being able to distinguish between male relatives based on Y-chromosomal marker analysis. In some embodiments, the rapidly-mutating Y-STR markers can be analyzed by a method employing multiplex PCR. Multiplex PCR can amplify 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 or all 13 of the rapidly-mutating Y-STR markers. In some embodiments, multiplex PCR can co-amplify additional Y-STR markers that are not part of the set of the subject rapidly-mutating Y-STR markers. In some embodiments, a multiplex PCR can provide for the co-amplification of one or

more autosomal STR markers, e.g. the CODIS STR markers, D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820, D16S539, TH01, TPDx, and CSF1 PO. Detailed descriptions for the development of multiplex PCR for STR analysis can be found, among other places in PCT patent application WO 2009/059049 A1. In some embodiments the PCR reactions are not multiplexed. The amplicons that are produced in non-multiplex PCR reactions can be combined prior to the analysis of an instrument, e.g. a fluorescent DNA fragment analyzer (such as an automated DNA sequencer) or a mass spectrometer. Mass spectroscopy of STR markers is described in, among other places, U.S. Pat. No. 6,090,558.

[0030] Other embodiments include sets of PCR primers for the co-amplification of at least two rapidly-mutating Y-STR markers. Embodiments include sets of PCR primers for the co-amplification of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 or all 13 of the rapidly-mutating Y-STR markers provided herein. In some embodiments, PCR primer sets can comprise primers for the co-amplification of Y-STR markers that are not rapidly-mutating Y-STR markers. In some embodiments, the set of PCR primers can comprise PCR primers for the co-amplification of STR markers present on an autosome.

[0031] The embodiments of the invention also include allelic ladders to aid in the identification of alleles of rapidly-mutating Y-STR markers. The allelic ladders can comprise sets of size standards for 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 or all 13 of the rapidly-mutating Y-STR markers. For each marker present in the allelic ladder, the allelic ladder can comprise standards for one or more alleles. An allelic ladder can comprise size standards for all known alleles of a given rapidly-mutating Y-STR marker, or any subset of known alleles. In some embodiments, the size standards in the allelic ladder can be labeled with one or more fluorescent dyes. In some embodiments an allelic ladder can further comprise size standards for autosomal STR markers. In some embodiments of allelic ladder can further comprise size standards for Y-STR markers that are not rapidly-mutating Y-STR markers.

[0032] Other embodiments of the subject invention include kits for the determination of the alleles for two or more rapidly-mutating Y-STR markers. Embodiments of the kits can comprise the subject sets of amplification primers. In some embodiments the kits can comprise one or more reagents used in nucleic acid amplification reactions. Examples of such reagents include, but are not limited to, DNA polymerases, dNTPs, buffers, nucleic acid purification reagents and the like. In some embodiments, the kits can comprise an allelic ladder designed to act as a size standard for the one or more rapidly-mutating Y-STR marker alleles generated (or potentially generated) by amplification primers present in the kit. Thus, in some embodiments, the kits can comprise allelic ladders specifically adapted to the amplicons generated by the use of the kit primers in an amplification reaction. For example a kit comprising primers for co-amplifying rapidly-mutating Y-STR markers DYF387S1, DYF399S1, and DYF404S1, can also include an allelic ladder having size standards for various alleles of rapidly-mutating Y-STR markers DYF387S1, DYF399S1, and DYF404S1. The kit can contain primers for co-amplifying all 13 RM-Y-STRs as well as an allelic ladder having appropriate size standards as would be known to one of skill in the art. The component size standards of an allelic ladder for given STR marker can be labeled with the same or different detectable labels, e.g., a

fluorescent dye, as are the primers used to generate the amplicons of the actual allele in the sample for analysis.

[0033] The invention may be better understood by reference to the following examples comprising experimental data. Such information is offered to be examples and is not intended to limit the scope of the claimed invention. Examples and data presented herein were published in K. Ballantyne, et al. "Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular Bases and Forensic Implications" *Am. J. Hum. Genet.* 87:341-353 (Sep. 10, 2010), and published online Sep. 2, 2010, each incorporated by reference herein.

EXAMPLES

DNA Samples

[0034] All father-son pairs used in the mutation rate study were confirmed in their paternity by molecular analyses, utilizing autosomal STRs, Y-STRs, HLA and RFLP genotyping and blood grouping, in addition to familial or governmental documentation. A threshold for paternity probability of 99.9% was set for inclusion in the study. Samples were obtained from the Berlin, Leipzig and Cologne areas of Germany, and the Warsaw and Wroclaw areas of Poland. Whole genome amplification using the GenomiPhi DNA Amplification kit (GE Healthcare, Little Chalfont, UK) was performed on the Leipzig samples due to low DNA quantities. WGA reactions were performed as recommended by the manufacturer, and products were purified using Invisorb 96 Filter Microplates (Invitex GmbH, Berlin, Germany). An additional set of independent samples from male relatives not used in the initial mutability screening from male families or pedigrees, used for verifying the value of identified rapidly mutating Y-STRs, came from the Greifswald, Kiel and Berlin areas of Germany, the Leuven area of Belgium, the Warsaw area of Poland, as well as Canada and Central Germany as described elsewhere 12. All families/pedigrees were confirmed by the same methods as the father-son pairs; pairs with complete genotypes for both the rapidly mutating (RM) Y-STRs and Yfiler Y-STRs were considered for analysis, or in the case of partial genotypes only those that showed a mutation at one or more loci were included. The use of all samples for the purpose of this study was in agreement with the institutional regulations and under informed consent.

Y-STR Markers and Genotyping Protocols

[0035] Y-STR markers were mostly selected from a previous study detailing a large number of 167 previously unknown Y-STRs 29, with the additional inclusion of Y-STRs known at the time of project commencement 42. The focus was on single-copy Y-STR markers in order to be able to fully confirm genotype differences by DNA sequence analysis when identifying mutations. However, given our aim to find RM Y-STRs, we included some additional multi-copy Y-STRs, especially those with high diversities (for which mutation confirmation was performed by independent genotyping). A complete list of loci, primer sequences and protocols can be found in the online Supplemental Data 51. Seventeen of the 186 Y-STRs were genotyped with a commercially available kit, the AmpF/STR Yfiler PCR Amplification kit (Applied Biosystems), following the manufacturer's instructions. Full descriptions of protocols and markers can be found in (28). The remaining 169 Y-STRs were genotyped using 54 multiplex assays including 1 to 5

markers each. PCRs were performed using three differing protocols, and details are provided in the Supplemental Data 51. In addition, 13 Y-STRs identified during the study as rapidly mutating (RM) Y-STRs were genotyped using three multiplex assays in an independent sample set of male relatives. All PCRs were performed on GeneAmp PCR System 9700 machines (Applied Biosystems) at the Department of Forensic Molecular Biology, Erasmus MC Rotterdam. Fragment length analysis was performed using the 3130x/Genetic Analyzer (Applied Biosystems) at Applied Biosystems, Foster City, USA. Profiles generated were genotyped using GeneMapper software (ID v 3.2, Applied Biosystems). Genotype differences were identified using in-house developed Microsoft Excel 2007 macros. All mutations were confirmed by DNA sequence analysis in Rotterdam of both the father and son at the Y-STR locus, as described in M. Goedbloed, et al. (2009) *Int. J. Legal. Med.* 123, 471-482. Multi-copy Y-STR loci with three or more alleles were not able to be sequenced, but mutations were confirmed by at least two independent fragment length analysis amplifications.

Statistical Data Analyses

[0036] Mutation rates for individual markers were estimated using a binomial hierarchical Bayesian model 43 using the Markov Chain Monte Carlo (MCMC) Gibbs sampling as implemented in WinBUGS, as described in Goedbloed. In brief, it was assumed that each mutation rate could be considered as a realization of the mutation rate underlying any Y-STR. In brief, we assumed that the mutation rate μ_i of Y-STR i was a sample from a common population distribution defined by hyperparameters. In that way, the estimated mutation rate of a Y-STR incorporates the information provided by the observed data on that Y-STR (number of observed mutations over all the observed father-son pair) and the information of the mutation rate of “the Y-STR” as estimated in the hyperparameter from all the Y-STRs. In practice, this implies that Y-STRs for which no mutation was observed are going to show a mutation rate (estimated from the posterior distribution) which is smaller than other Y-STRs where a large number of mutations are observed, but is always different from 0.

[0037] The mutation rate of each Y-STR was coded in a logit form, and assumed to follow a normal distribution with parameters and τ_i to be estimated, as well as the particular mutation rates of each STR. As only very limited data was available prior to our study for the range of Y-STR mutation rates, we assumed diffuse, non-informative prior distributions for the hyperparameters. A non-informative prior normal distribution ($\mu=0$, $\sigma^2=1 \times 10^{-6}$) was specified for the hyperparameter and a prior diffuse gamma distribution with parameters $\alpha=1 \times 10^{-5}$ and $\beta=1 \times 10^{-5}$ for the parameter. Three MCMC chains using the Gibbs sampler were generated in parallel when estimating the mutation rate for each locus, with 100,000 runs performed for each chain. Mean, median and 95% credible intervals (CI) were estimated from the three chains after discarding the first 50,000 runs and performing a thinning of 15 in order to reduce the amount of autocorrelation between adjacent simulations. Locus-specific differences in mutation rates between the sampling populations (Cologne, Berlin, Leipzig, Warsaw, Wrocław) were tested by means of a permutation analysis. The average mutation rate for each locus and each population was compared to a hypothetical permuted population, where each father-son pair had been assigned to a population at random, maintaining

the original sample sizes for each locus. The number of times the permuted averaged mutation rate was larger than the observed rate was recorded, and used to obtain the one tail p value over 100,000 iterations. The lack of significant differences between populations allowed pooling of mutation rates across populations.

[0038] In order to investigate the mutation rate of the Yfiler and RMY-STR sets rather than of each marker within the set, the total number of mutations observed between each father-son pair for each set was computed, given the number of Y-STRs analyzed. This parameter was then modeled under the Bayesian paradigm with a Poisson distribution. A prior with a Gamma distribution was used with a diffuse shape of 1 and a scale of 200, implying a mutation rate with a mean of 0.005 and a variance of 40000. The posterior distribution followed a conjugate Gamma distribution with shape of $1+(\text{total number of mutations})$ and scale of $1/(200+\text{total number of markers used})$. In order to estimate the probability of observing at least one mutation in each set, 100,000 Monte Carlo replicates were performed with the rgamma function of the R package 45 from the estimated shape and scale of the posterior distribution of each set of Y-STRs.

[0039] For the RM Y-STR set a median mutation rate of 0.0197 (95% credible interval 0.018–0.022) was estimated that is about 7-fold higher as revealed for the Yfiler set consisting of 17 markers with a median rate of 0.0028 (95% credible interval ranging from 0.0023 to 0.0035). Next, the probability of observing at least one mutation per Y-STR set in a given father-son pair, reflecting the minimal criteria for differentiating male relatives, was estimated as 1 minus the probability of observing 0 mutations, which is directly estimated from a Poisson distribution: The probability of observing at least one mutation (k) within either of the YSTR sets in any given father-son pair was directly estimated from the Poisson distribution:

$$P(k>0)=1-P(k=0)=1-e^{-Nm},$$

with N representing the number of markers and m representing the average mutation rate of the set of markers obtained from the sampling from the posterior distribution. Assuming that all Y-STRs per set have been genotyped successfully, and using the posterior estimates of the mutation rate for each set of markers, the probability of observing at least one mutation with the RM Y-STR set is 0.1952 (95% credible interval of 0.177 to 0.21). This value is more than four times higher than that estimated for the Yfiler set with 0.047 (95% credible interval of 0.038 to 0.057), although six more markers are included in the Yfiler set relative to the RMY-STR set. The molecular factors determining mutation rates were modeled using a Poisson regression with in-house developed Matlab scripts (v7.6.0.324, The Mathworks, Inc., Natick, Mass., USA). The mutation rate was modeled as a function dependent on of the repeat length, the sequence motif, the complexity of the locus and the length of the repeat in base pairs (tri-, tetra-, penta- or hexanucleotide), as:

$$p(y | \theta) = \prod_{i=1}^n \frac{1}{y_i!} (x_i \theta)^{y_i} e^{-x_i \theta}$$

where θ is assumed to be dependent on the factors described above, in the form

$$\theta = e^{\alpha L + \beta S + \gamma C + \delta V + \epsilon R + \zeta N}$$

where L represents the length of the allele (number or repeats, either of the longest homogenous array or the total locus), S represents the sequence motif (comprised of the number of A,T,C or G nucleotides in the repeated sequence motif), C represents the complexity of the locus, either in binary or quantitative form, V is the number of variable motifs present, R is the repeat length, and N is the copy number of the locus. A stepwise regression procedure was used, with probability to enter ≤ 0.05 , probability to remove ≤ 0.10 . For clarity, the methods used for defining and calculating the number of repeats within a locus, and the complexity of that locus, are elucidated below.

[0040] Locus designations were modeled after Kayser et al., where at least 3 consecutive repeats of the same motif are required to define a given repeat segment as a locus, and any interruption of more than 1 base, but less than a full unit, is classed as ending the locus. Individual Y-STR loci contained between 1 and 5 repeat blocks, as in, for example, DYS612 with 5 blocks (CCT)5(CTT)1(TCT)4(CCT)1(TCT)19. If a locus contained more than one variable segment, and repeat numbers could not be assigned to all individuals at all repeat segments accurately, the locus was removed from the regression analysis. A segment was defined as variable if a variation in repeat number was seen in any individual sequenced, relative to the remainder of the population.

[0041] Number of repeats: The number of repeats in the longest homogenous array was directly counted, and the population average calculated for each locus. In addition, any additional repeats around the longest array were added to calculate the total number of repeats for each locus. In the above example for DYS612, the length of the longest array is 19, while the total number of repeats is 30.

[0042] Repeat Length: The length in base pairs of the repetitive motif, which ranged from 3 to 6 (included tri-, tetra-, penta-, hexa- and heptanucleotide repeats).

[0043] Complexity: Two complexity statistics were calculated per locus. First, a binary classification system was used, where loci with only one repetitive segment (e.g. (GATA)10) were classified as simple, while any locus with two or more repetitive segments consisting of more than three consecutive repeats (e.g. (GATA)10(CATA)3) was complex. Second, more quantitative information was provided by Kayser et al.'s complexity formula:

$$C = \frac{n^2}{(n-1)^2} \left(1 - \sum_{i=1}^m \left(\frac{s_i}{n} \right)^2 \right) \left(1 - \sum_{i=1}^l \left(\frac{b_i}{n} \right)^2 \right)$$

where n is the total number of repeats in the locus, s_i is the number of repeats of the ith sequence motif, and b_i is the number of repeats in the ith block. Correlation and log linear regression analyses were carried out in SPSS v15.0 (SPSS Inc.), as were all mean comparison tests (utilizing ANOVA, Mann-Whitney U and Kruskal Wallis).

[0044] Repeat Length: The length in base pairs of the repetitive motif, which ranged from 3 to 6 (included tri-, tetra-, penta-, hexa- and heptanucleotide repeats).

Mutation Rates of Y-STR Markers

[0045] In order to define the expectation for a given RM Y-STR set to differentiate between male relatives, and to compare such potential with that of the commonly-used

YFiler set, an average mutation rate for each of the two Y-STR sets applying a Bayesian approach was obtained. The number of mutations observed in one father-son pair for a set of STRs was modeled by means of a Poisson distribution. A prior conjugate Gamma distribution with a diffuse shape of 1 and a scale of $1/0.005$ was used. The posterior distribution followed a Gamma distribution with shape of $1 + \text{total number of mutations}$ and scale of $1/(1/0.005 + \text{total number of markers used})$ was obtained and 100000 Monte Carlo replicates were performed.

[0046] Furthermore, to test in independent samples whether the new RM Y-STR set is practical and useful for differentiating male relatives, genotyping was performed on both marker sets in 107 pairs from 80 male pedigrees who were related by between 1 and 20 generations within their pedigrees and compared the findings with those from YFiler also generated. Pedigrees came from the Greifswald and Kiel (N. von Wurmb-Schwark, V. Mályusz, E. Simeoni, E. Lignitz, M. Poetsch, *For. Sci. Int.* 159, 92-97 (2006)), as well as Berlin (new to this study) areas of Germany, the Leuven area of Belgium (new to this study), the Warsaw area of Poland (new to this study), as well as from Canada C. Moreau, H. Vézina, V. Yotova, R. Hamon, P. de Kniff et al., *Am. J. Phys. Anthropol.* 139, 512-522 (2009), M. Vermeulen, A. Wollstein, K. van der Gaag, O. Lao, Y. Xue et al., *For. Sci. Int. Genet.*, 3, 205-213 (2009) and Central Germany M. Kayser, M. Vermeulen, H. Knoblach, H. Schuster, M. Krawczak, L. Roewer, *For. Sci. Int. Genet.* 1, 125-128 (2007)), as described elsewhere. All pedigrees were confirmed by DNA data (including autosomal STR, HLA and RFLP typing, Y-STR and Y-SNP typing, and mtDNA sequencing amongst various pedigrees), as well as additionally by familial or governmental documentation records. Only pairs which had complete genotypes for both sets, or in the case of partial genotypes, showed a mutation at one or more loci, were included in the calculations. Results are provided in FIG. 2. The RM Y-STR set distinguished over 65% of pairs by at least 1 mutation, reflecting a 5-fold increase in the level of male relative differentiation compared to the YFiler set with only 13%, similar to our statistical expectations from the initial father-son pair analyses. Within the pedigrees, the RM Y-STR set distinguished 60% of father-son pairs, 54% of brothers, and 87% of second cousins. If relatives were separated by more than 11 meioses, 100% of individuals were separated by 1 or more mutations using the RM Y-STR set. In contrast, the Y-filer set distinguished in this dataset no father son pairs, no second cousins, and only 6% of brothers in this dataset.

[0047] 186 tri-, tetra- penta- and hexanucleotide Y-STR markers were screened for mutations in up to 1966 DNA-confirmed father-son pairs per marker by multiplex fluorescence-based fragment length analysis, giving direct observation of 352,999 meiotic transfers (for technical details see Table 1). To confirm mutations, all Y-STR genotype differences observed between fathers and their sons were confirmed by DNA sequence analysis for single copy and duplicated markers, or by duplicate fragment length genotyping analysis for multi-copy Y-STRs with more than 2 copies (where sequence analysis was not informative). Overall, we identified 924 confirmed mutations at 120 (64.5%) of the 186 Y-STR markers studied (details of each mutation observed can be found in Supplemental Data S2). For 66 Y-STR markers, the up to 1966 father-son pairs analyzed did not allow us to detect mutations due to a very low underlying mutation rate. The large number of Y-STR markers employed identified the range of Bayesian-based mutation rates estimated from

the median of the posterior distribution to be between 3.81×10^{-4} (95% CI 1.38×10^{-5} to 2.02×10^{-3}) and 7.73×10^{-2} (6.51×10^{-2} to 9.09×10^{-2}) per marker, per generation (FIG. 1, Table 1). Ninety-one Y-STR markers (48.9%) had mutation rates in the order of 10^{-3} , a further 82 markers (44%) in the order of 10^{-4} , and 13 (6.9%) in the order of 10^{-2} . Across all 186 Y-STR markers, the average mutation rate was 3.35×10^{-3} (95% CI 1.79×10^{-3} to 6.38×10^{-3}) with an average rate of 4.26×10^{-3} (95% CI 2.38×10^{-3} to 7.60×10^{-3}) for the 122 tetranucleotide repeats as the largest repeat-length subgroup of Y-STR markers included here. Notably, the 13 Y-STR markers with mutation rates above 1×10^{-2} representing only 7% of the markers studied, which we termed “Rapidly mutating Y-STRs” (RM Y-STRs), covered a large number of 462 of the 924 (50%) mutations observed in the study.

[0048] Number of repeats. Two estimates of the average number of repeats were calculated for each Y-STR locus i) the average repeat number in the longest homogenous array; and ii) the repeat number of the longest homogeneous array plus any non-variable repeats immediately adjacent (in accordance with previously defined rules for motif structure 29). Our regression analysis showed that while the number of repeats in the longest homogenous array did influence the mutation rate significantly, with higher numbers of repeats increasing the mutation rate (Wald $X^2=2.41 \times 10^6$, $p<0.0001$), including the number of non-variable repeats surrounding the array provided slightly more accurate information to the model (Wald $X^2=3.03 \times 10^6$, $p<0.0001$, FIG. 2). The effect size within the model was estimated with a partial η^2 of 0.798, indicating that the variance in the total number of repeats between loci accounts for ~78% of the overall (effect+error) variation in Y-STR mutation rates observed. In addition, a statistically significant exponential relationship was observed between the total number of repeats and the allele-specific mutation rate ($R^2=0.707$, $p=6.84 \times 10^{-9}$). In addition, there was a strong relationship between the total number of repeats and the direction of mutation (FIG. 3). Longer alleles displayed an exponential and statistically significant tendency towards repeat losses (contractions) ($R^2=0.585$, $p=8.27 \times 10^{-7}$), while shorter alleles gained repeats (expansion) significantly more frequently ($R^2=0.238$, $p=0.011$). The expansion mutation rate had a quadratic distribution, with a vertex around 19 repeats.

Male Relative Differentiation by RM Y-STRs

[0049] We identified 13 rapidly-mutating (RM) Y-STR markers (all with mutation rates $>1 \times 10^{-2}$); DYS387S1, DYS399S1, DYS403S1, DYS404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627 (FIG. 1, Table 1). Four of these 13 RM Y-STR markers are multi-copy systems (DYS387S1 with two, DYS399S1 with three, DYS403S1 with four, DYS404S1 with two and DYS526 with two copies), whereas nine were single-copy Y-STR markers (although six of these markers contained multiple Y-STR loci within the single amplicon, and only two, DYS570 and DYS576, were simple repeats with only one Y-STR locus respectively). The 13 RM Y-STRs were combined into a set under the hypothesis that closely related males (even father-son or brother pairs) may be differentiable by Y-STR mutations if RM Y-STRs are combined. In principle, one mutation at one of the 13 RM Y-STRs would be enough for individual differentiation.

[0050] In order to define a statistical expectation for the RM Y-STR set to differentiate between male relatives, and to

compare their potential with that of the commonly used Yfiler set, we first computed the mutation rate observed for each of the two Y-STR sets by means of a Bayesian approach. The number of mutations observed in each father-son pair for each set of Y-STRs was modeled by means of a Poisson distribution. For the RM Y-STRs a median mutation rate of 1.97×10^{-2} (95% CI 1.8×10^{-2} – 2.2×10^{-2}) of the posterior distribution was estimated, which was 6.5-fold higher than that estimated for Yfiler Y-STRs with a median rate of 3.0×10^{-3} (95% CI ranging from 2.39×10^{-3} to 3.72×10^{-3}). Next, the probability of observing at least one mutation in each of the two Y-STR sets for a given father-son pair was estimated, reflecting the minimal criteria for differentiating male relatives. Assuming that all Y-STRs per set were genotyped successfully, and using the posterior estimates of the mutation rate for each set of Y-STR markers, the probability of observing at least one mutation with the RM Y-STR set was 0.1952 (95% CI of 0.177 to 0.21). This value was surprisingly more than four times higher than that estimated for the Yfiler set with 0.047 (95% CI of 0.038 to 0.057). The probability of observing at least one mutation with the RM Y-STR set was statistically significantly higher than for the Yfiler set ($p<5.0 \times 10^{-07}$). Finally, samples were empirically tested independent of those samples used for mutation rate establishment whether the new RM Y-STR set is practically useful for differentiating male relatives. For this, 103 male relative pairs from 80 male pedigrees who were related by between 1 and 20 generations within their pedigrees were genotyped and compared with the findings with those obtained from Yfiler kit in the same samples. Overall, the RM Y-STR set distinguished 70.9% pairs of male relatives by at least 1 mutation, reflecting a 5-fold increase in the level of male relative differentiation compared to the Yfiler kit set with only 13%; notably, the significant difference ($t=6.389$, $p<0.0001$) is similar to statistical expectations from the initial father-son pair analyses (FIG. 4 and Table 3). Within the pedigrees, the RM Y-STR set distinguished 70% of father-son pairs, 56% of brothers, and 67% of cousins (FIG. 4 and Table 3). In contrast, the Yfiler set was not able to differentiate any of the father-son pairs nor cousins, and only 6% of the brothers in this dataset (FIG. 4 and Table 3). Furthermore, all relatives separated by more than 11 generations were differentiable by 1 or more mutations using the RM Y-STR set, but only 33% with the Yfiler set.

[0051] All of the compositions and methods disclosed and claimed herein can be made and executed without undue experimentation in light of the present disclosure. While the compositions and methods of this invention may have been described in terms of specific examples or preferred embodiments, these examples and embodiments are in no way intended to limit the scope of the claims, and it will be apparent to those of skill in the art that variations may be applied to the compositions and/or methods and in the steps or in the sequence of steps of the methods described herein without departing from the concept, spirit and scope of the invention. More specifically, it will be apparent that certain agents which are both chemically and physiologically related may be substituted for the agents described herein while the same or similar results would be achieved. All such similar substitutes and modifications apparent to those skilled in the art are deemed to be within the spirit, scope and concept of the invention as defined by the appended claims.

SEQUENCE LISTING

The patent application contains a lengthy "Sequence Listing" section. A copy of the "Sequence Listing" is available in electronic form from the USPTO web site (<http://seqdata.uspto.gov/?pageRequest=docDetail&DocID=US20110263437A1>). An electronic copy of the "Sequence Listing" will also be available from the USPTO upon request and payment of the fee set forth in 37 CFR 1.19(b)(3).

What is claimed is:

1. A set of amplification primer pairs, comprising primers for the amplification of at least 2 Y-STR markers selected from the group consisting of DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627.

2. A set of primers according to claim 1, wherein the primers can be used to co-amplify at least 3-13 loci from the group.

3. A set of primer according to claim 1, wherein the primers can be used to amplify all loci from the group.

4. A method of identifying an individual, the method comprising determining the allele of at least 2 Y-STR markers selected from the group consisting of DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627.

5. The method of claim 4, wherein the allele is identified by PCR.

6. The method of claim 5, wherein the PCR is multiplex PCR that co-amplifies the at least 3 of the markers.

7. The method of claim 5, wherein the PCR uses primers that are labeled with a fluorescent dye.

8. The method of claim 4, wherein the allele is identified by mass spectroscopy, capillary electrophoresis, or gel electrophoresis.

9. The method of claim 4, wherein the PCR co-amplifies at least one the loci and an autosomal STR.

10. The method of claim 9, wherein the autosomal STR is selected from the group consisting of D3S1358, vWA, FGA, D8S1179, D21 S11, D18S51, D5S818, D13S317, D7S820, D16S539, THO1, TPDx, and CSF1 PO.

11. A kit for identifying the allele of at least 2 Y chromosome STRS markers, wherein the markers are selected from the group consisting of DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627, the kit comprising primers for the amplification of at least 3 loci, and an allelic ladder representative of the selected markers.

12. An allelic ladder size standard for calling one or more alleles of an STR from at least 2 of the Y-STR markers selected from the group consisting of DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627.

13. A set of amplification primer pairs, comprising at least one primer pair selected from the group consisting of DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627.

14. A set of amplification primer pairs of the identification of a male, comprising primers for the Y-STR markers consisting of DYF387S1, DYF399S1, DYF403S1, DYF404S1, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627.

* * * * *