

(12) STANDARD PATENT
(19) AUSTRALIAN PATENT OFFICE

(11) Application No. **AU 2010329552 B2**

(54) Title
Categorization of DNA samples

(51) International Patent Classification(s)
C12Q 1/68 (2006.01)

(21) Application No: **2010329552**

(22) Date of Filing: **2010.12.08**

(87) WIPO No: **WO11/070441**

(30) Priority Data

(31) Number	(32) Date	(33) Country
61/306,201	2010.02.19	US
61/325,977	2010.04.20	US
61/285,758	2009.12.11	US

(43) Publication Date: **2011.06.16**

(44) Accepted Journal Date: **2015.11.26**

(71) Applicant(s)
Nucleix Ltd

(72) Inventor(s)
Wasserstrom, Adam;Frumkin, Dan

(74) Agent / Attorney
Knightsbridge Patent Attorneys, Gpo Box 4691, Melbourne, VIC, 3001

(56) Related Art
SAKAMOTO, H., et al., Genes to Cells, 2007, vol. 12, pages 1123-1132
EP 1 213 360
EP 1 748 080
WO 2009/083989
OAKES, C. C., et al., Epigenetics, 2006, vol. 1, pages 146-152 *
US 2009/0305234

(43) International Publication Date
16 June 2011 (16.06.2011)(10) International Publication Number
WO 2011/070441 A2(51) International Patent Classification:
C12Q 1/68 (2006.01)(21) International Application Number:
PCT/IB2010/003397(22) International Filing Date:
8 December 2010 (08.12.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/285,758 11 December 2009 (11.12.2009) US
61/306,201 19 February 2010 (19.02.2010) US
61/325,977 20 April 2010 (20.04.2010) US

(72) Inventor; and

(71) Applicant : WASSERSTROM, Adam [IL/IL]; 1 Haogen St., Apt. 14, 74100 Nes-Ziona (IL).

(72) Inventor; and

(75) Inventor/Applicant (for US only): FRUMKIN, Dan [IL/IL]; 4 Reiness St., Apt. 13, 61999 Tel-aviv (IL).

(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, 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, PE, PG, PH, PL, PT, RO, RS, RU, 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, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, 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).

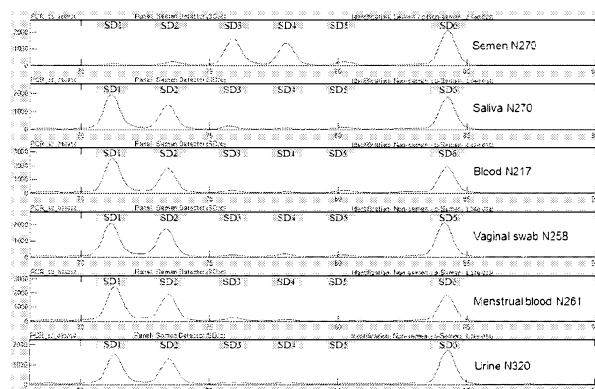
Published:

— without international search report and to be republished upon receipt of that report (Rule 48.2(g))

— with sequence listing part of description (Rule 5.2(a))

(54) Title: CATEGORIZATION OF DNA SAMPLES

Figure 12



(57) **Abstract:** The present application describes methods for accurate and cost-effective categorization of DNA samples into different types of in vitro generated DNA or different types of natural DNA such as from different tissues and/or physiological/pathological states. The invention achieves categorization by comparing "signal ratios" that are correlated to ratios of methylation levels at specific genomic loci, but does not rely on calculation of actual methylation levels at any genomic locus. Therefore the disclosed inventive method eliminates the requirement for external DNA species and controls, thereby simplifying and increasing the accuracy of the assay. The described inventive technology also enables performing the categorization of DNA together with DNA profiling in the same reaction, thereby allowing for concomitant categorization and determination of identity of the samples.

CATEGORIZATION OF DNA SAMPLES

This International Application claims priority to United States Provisional Applications 61/285,758 filed on December 11, 2009, 61/306,201 filed on February 19, 2010, and 61/325,977 filed on April 20, 2010, which are each all incorporated herein by reference in their entirety.

FIELD OF THE INVENTION

The present inventive technology relates to methods for accurate and cost-effective categorization and profiling of DNA samples to identify the originating source and identity of the DNA sample or mixed DNA samples.

SUMMARY OF THE INVENTION

The present invention relates to methods for accurate and cost-effective categorization of DNA samples into different types of *in vitro* generated DNA or different types of natural DNA such as from different tissues and/or physiological/pathological states. The invention achieves categorization by comparing “signal ratios” that are correlated to ratios of methylation levels at specific genomic loci, but does not rely on calculation of actual methylation levels at any genomic locus. Therefore the invention setup eliminates the requirement for external DNA species and/or controls, thereby both simplifying and increasing the accuracy of the assay. The inventive technology also enables performing the categorization of DNA together with DNA profiling in the same reaction, thereby allowing for concomitant categorization and determination of identity of the samples.

One aspect of the present invention is a method for categorization of a DNA sample, the method comprising:

- (A) digesting a DNA sample with a methylation-sensitive and/or methylation-dependent restriction endonuclease;
- (B) performing PCR on the digested DNA amplifying at least two genomic loci of which at least one is a restriction locus;
- (C) determining the intensity of the signal of each amplification product;

(D) calculating “signal ratios” between the intensities of the signals produced by the loci; and

(E) comparing the signal ratios to reference values corresponding to different categories of DNA, wherein the category whose reference values correspond best to the signal ratios is determined to be the category of the DNA sample. That is, for instance, the approximation of the value of a signal ratio to the value of a reference value of a particular DNA category indicates the categorical source of the DNA sample. So if a signal ratio that has a value that is close to the value of a reference category, then the likelihood that the tested DNA sample originated from that particular type of category – based on the similarity in values between the signal ratio and the reference value – is high. Thus, the signal ratio(s) that best correspond to, or best approximate, a particular reference value(s) for a particular DNA category makes it likely that that DNA sample originated from that particular DNA categorical source. For examples about how to compare the signal ratios and category reference values see the disclosure elsewhere in this application, such as in the section entitled Algorithm and Software, which explains how to calculate probability distributions and likelihood probabilities that a test DNA sample can be categorized to one or more reference categories.

In one embodiment, DNA digestion and PCR are performed in a single biochemical reaction in which the DNA template, digestion and amplification enzymes, buffers, primers, and accessory ingredients are placed together in a test tube and then the test tube is closed and placed in a thermal cycler, where the reaction takes place.

In one embodiment, the methylation-sensitive restriction endonuclease is unable to cut or digest a DNA if its recognition sequence is methylated. In one embodiment, a methylation-sensitive restriction endonuclease is selected from the group consisting of AatII, Acc65I, AccI, AciI, ACII, AfeI, AgeI, ApaI, ApaLI, AscI, AsiSI, AvaI, AvaII, BaeI, BanI, BbeI, BceAI, BcgI, BfuCI, BglI, BmgBI, BsaAI, BsaBI, BsaHI, BsaI, BseYI, BsiEI, BsiWI, BslI, BsmAI, BsmBI, BsmFI, BspDI, BsrBI, BsrFI, BssHII, BssKI, BstAPI, BstBI, BstUI, BstZ17I, Cac8I, ClaI, DpnI, DrdI, EaeI, EagI, EagI-HF, EciI, EcoRI, EcoRI-HF, FauI, Fnu4HI, FseI, FspI, HaeII, HgaI, HhaI, HincII, HincII, HinfI, HinP1I, HpaI, HpaII, Hpy166ii, Hpy188iii, Hpy99I, HpyCH4IV, KasI, MluI, MmeI, MspAII, MwoI, NaeI, NarI, NgoNIV, Nhe-HFI, NheI, NlaIV, NotI, NotI-HF, NruI, Nt.BbvCI, Nt.BsmAI, Nt.CviPII, PaeR7I, PleI,

PmeI, PmlI, PshAI, PspOMI, PvuI, RsaI, RsrII, SacII, SalI, SalI-HF, Sau3AI, Sau96I, ScrFI, SfiI, SfoI, SgrAI, SmaI, SnaBI, TfiI, TscI, TseI, TspMI, and ZraI.

In another embodiment, a methylation dependent restriction enzyme can be used to digest a DNA sample. A methylation dependent restriction endonuclease digests only methylated DNA. Examples of such restriction enzymes include but are not limited to McrBC, McrA, and MrrA.

In one embodiment, at least one restriction locus is known to be methylated in natural DNA.

In another embodiment, at least one restriction locus is known to be unmethylated in natural DNA.

In another embodiment, at least two restriction loci are known to be differentially methylated in natural DNA.

In another embodiment, the signal ratio between at least two loci is known to be different for at least two potential categories.

In another embodiment, primer pairs used to amplify loci are selected to amplify the core short tandem repeat (STR) loci identified in the Federal Bureau of Investigation's Combined DNA Index System (CODIS).

In one embodiment, the CODIS loci are human genomic loci selected from the group consisting of D16S539 (SEQ ID NO. 1), D7S820 (SEQ ID NO. 2), D13S317 (SEQ ID NO. 3), D5S818 (SEQ ID NO. 4), CSF1PO (SEQ ID NO. 5), TPOX (SEQ ID NO. 6), TH01 (SEQ ID NO. 7), vWA (SEQ ID NO. 8), FGA (SEQ ID NO. 9), D21S11 (SEQ ID NO. 10), D8S1179 (SEQ ID NO. 11), D18S51 (SEQ ID NO. 12), and D3S1358 (SEQ ID NO. 13).

In a particular embodiment, the forward and reverse primers of a restriction and/or profile locus are designed to anneal to SEQ ID NO. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, or 25.

In one embodiment, the forward and reverse primers of a locus are available in a Promega Corporation commercial kit selected from the group consisting of PowerPlex® 16 HS (Cat.# DC2100, DC2101), PowerPlex® 16 (Cat.# DC6530, DC6531), PowerPlex® 2.1

(Cat.# DC6470, DC6471), PowerPlex® 16 BIO (Cat.# DC6540, DC6541), and PowerPlex® ES Systems (Cat.# DC6730, DC6731).

In one embodiment, the forward and reverse primers of a locus are available in an Applied Biosystems/Life commercial kit selected from the group consisting of SGM, SGM+, AmpFISTR Identifiler, AmpFISTR Profiler, AmpFISTR ProfilerPlus, AmpFISTR ProfilerPlusID, AmpFISTR SEfiler, AmpFISTR SEfiler Plus, AmpFISTR Cofiler, AmpFISTR Identifiler Direct, AmpFISTR Identifiler Plus, AmpFISTR NGM, AmpFISTR Y-filer, AmpFISTR Minifiler.

In one embodiment, the forward and reverse primers of a locus are available in a Qiagen commercial kit selected from the group consisting of Investigator ESSPlex, Investigator ESSPlex SE, Investigator Nonaplex ESSPlex, Investigator Hexaplex ESSPlex, Investigator Triplex AFS QS, Investigator Triplex DSF, Investigator IDplex, Investigator Decaplex SE, Investigator HDplex, Investigator Argus X-12, Investigator Y-12 QS, Investigator DIPplex.

In another embodiment, the restriction loci contain a HpaII recognition sequence and are selected from the group consisting of ADD6, ADD10, ADD17 (SEQ ID NOs. 26-28) and Hypo23, Hypo28, Hypo33 (SEQ ID NOs. 29-31). In one embodiment, the DNA sample is mammalian DNA or plant DNA. In one embodiment, the mammalian DNA is DNA from a mammal selected from the group consisting of human, ape, monkey, rat, mouse, rabbit, cow, pig, sheep, and horse. In a particular embodiment, the mammalian DNA is human DNA. In one embodiment, the human DNA is from a male. In another embodiment, the human DNA is from a female.

In one embodiment, the human DNA sample is obtained from a male or female who has a disorder characterized by abnormal DNA methylation. In one embodiment, the disorder is a neurodevelopmental disorder of ICF (immunodeficiency, centromeric instability, and facial anomalies), Rett syndrome, and fragile X syndrome.

In one embodiment, the plant DNA is DNA from a plant. In one embodiment, the plant is a monocotyledonous plant selected from the group consisting of wheat, turf grass, cereal, maize, rice, oat, barley, sorghum, orchid, iris, lily, onion, banana, sugarcane, sorghum, and palm. In another embodiment, the plant is a dicotyledonous plant selected from the group consisting of avocado, potato, tobacco, tomato, sugarbeet, broccoli, cassava, sweet

potato, pepper, cotton, poinsettia, legumes, alfalfa, soybean, carrot, strawberry, lettuce, oak, maple, walnut, rose, mint, squash, daisy, and cactus.

In one embodiment of the method, at least two loci are amplified by PCR with fluorescently labeled primers and wherein pairwise comparisons of the intensity of the fluorescent signals corresponding to the loci is performed in order to calculate signal ratios for each pair of loci.

In one embodiment, primer pairs are used to amplify pairs of loci selected from the group of pairs consisting of D3S1358/D18S51, D3S1358/D7S820, D3S1358/Penta_D, D3S1358/TPOX, D3S1358/FGA, TH01/Penta_D, D21S11/D18S51, D21S11/D7S820, D21S11/Penta_D, D21S11/AMEL, D21S11/TPOX, D21S11/FGA, D5S818/D18S51, vWA/D18S51, D5S818/Penta_E, vWA/Penta_E, D5S818/D7S820, D5S818/Penta_D, D5S818/TPOX, D5S818/FGA, D13S317/D7S820, D13S317/Penta_D, D13S317/TPOX, D13S317/FGA, D16S539/D7S820, CSF1PO/D7S820, vWA/D7S820, D8S1179/D7S820, D16S539/TPOX, D16S539/FGA, CSF1PO/Penta_D, CSF1PO/TPOX, vWA/Penta_D, AMEL/TPOX, AMEL/FGA, vWA/D8S1179, vWA/TPOX, vWA/FGA, D8S1179/TPOX, D8S1179/FGA.

In another embodiment, the restriction loci amplicons are smaller than the smallest amplicons used in DNA profiling, which is about 100 bps in size.

In one aspect of the present inventive technology, quantification of amplicon signals is performed and not necessarily the quantification of the amount of DNA templates. By calculating ratios of amplicon signals, no standard curve or reference DNA is needed since it is unnecessary to calculate actual DNA concentrations. In this embodiment, the ratios of amplicon signals is calculated. The signals may be signals detected from any kind of labeling used for detecting PCR amplification products, such as, but not limited to, ethidium bromide staining, sybr green staining, silver staining, or by fluorescence.

With respect to the latter, in one embodiment, the PCR reaction is performed with fluorescently labeled primers, and determination of the intensity of the signals of the amplification products is achieved by separation of products by capillary electrophoresis and quantification of fluorescence signals.

In another embodiment, the PCR is real time PCR, and determination of the intensity of the signals of the amplification products is achieved by quantification of fluorescence signals.

In a particular embodiment, the methylation-sensitive restriction endonuclease is HhaI.

In a particular embodiment, the methylation-sensitive restriction endonuclease is HhaI and the first restriction locus is the TPOX locus.

In a particular embodiment, the methylation-sensitive restriction endonuclease is HhaI and the first restriction locus is the FGA locus.

In a particular embodiment in which categorization is according to semen or non-semen, the methylation-sensitive restriction endonuclease is HhaI, the restriction loci are SD1, SD2, SD3, and SD4 (SEQ ID NOs. 33-36), a digested control locus is SD5 (SEQ ID NO. 37), and an undigested control locus is SD6 (SEQ ID NO. 38).

In a particular embodiment in which categorization is according to blood, saliva, semen, or skin epidermis, the methylation-sensitive restriction endonuclease is HhaI and the restriction loci are SEQ ID NOs. 39-53.

In a particular embodiment in which categorization is performed together with DNA profiling, and is according to semen or non-semen, the methylation-sensitive restriction endonuclease is HhaI, and the restriction loci are SEQ ID NO. 40 and SEQ ID NO. 54.

In one embodiment, a pair of loci are selected from loci with known methylation levels in at least two DNA categories, such that one of the loci in the pair is known to have a higher methylation level than the second locus in that pair in one DNA category but not in the other.

In another embodiment, a pair of loci are selected without knowledge of methylation levels as follows: select a pair of random genomic loci; design fluorescently labeled primers for amplifying each locus; amplify both loci in DNA samples obtained from different categories; separate amplicons with capillary electrophoresis; calculate signal ratios in the different DNA samples; if ratios calculated from samples corresponding to at least one category are significantly different (e.g. as determined by the Kolmogorov-Smirnov test with

a threshold of $p = 0.05$) than those obtained from samples corresponding to at least one other category, select the pair for use in assay. In one embodiment, the calculated ratios are compared to a set of ratios obtained from natural samples and to a set of ratios obtained from artificial samples.

With respect to the latter, in one embodiment, the respective mean signal ratios of natural and artificial DNA samples digested with HhaI for specific pairs of loci determined herein are shown below.

	Natural DNA	Artificial DNA
FGA(PowerPlex16)/Hypo12	10.3	0.89
FGA(PowerPlex16)/Hypo28	12.4	0.76
TPOX(PowerPlex16)/Hypo12	9.7	1.12

In another embodiment, the calculated ratios are compared to a set of ratios obtained from DNA samples extracted from different tissue categories. See Tables 1-4 presented in Figures 7-10 for additional examples of loci sequences and illustrative signal ratios and results used and obtained in accordance with the inventive categorization methods disclosed herein.

In one embodiment this method, or any method described herein, further comprises simultaneously profiling a DNA sample and determining whether the category of the DNA. In one embodiment, the step of profiling a DNA sample is performed according to instructions provided in a commercially available DNA profiling kit. In another embodiment, the commercially available DNA profiling kit is the PowerPlex®16 profiling kit (Promega Corporation).

In one embodiment, with respect to the latter, categorization of DNA is according to tissue types. Any tissue type can be used such as, but not limited to blood, saliva, semen, epidermis, urine, plasma, and hair.

Accordingly, one aspect of the present invention is a method for categorizing a DNA sample, comprising:

- (A) digesting a DNA sample with at least one of a methylation-sensitive restriction endonuclease and a methylation-dependent restriction endonuclease;
- (B) amplifying at least two genomic loci from the digested DNA, wherein at least one of the loci is a restriction locus;
- (C) determining the signal intensity of each amplification product;
- (D) calculating signal ratios between the signal intensities of the amplification products; and
- (E) comparing the signal ratios to reference values of different DNA categories,

wherein the approximation of the value of a signal ratio to the values of reference ratios of a particular DNA category indicates the categorical source of the DNA sample.

In one embodiment, the reference values against which the signal ratios are compared are from at least one DNA category selected from the group consisting of tissue type, cell type, physiological condition, pathological condition, age, ethnicity, sex, methylation level, species, cell lines, natural DNA and artificially-synthesized DNA, risk of developing a pathological condition, prenatal conditions, prognosis, propensity to respond to treatment, effects of cell line sub-culturing, response to medication.

In another embodiment, the tissue type or cell type is selected from the group consisting of blood, saliva, semen, epidermis, urine, plasma, hair, menstrual blood, vaginal cells and/or secretion, sweat, feces, brain, esophagus, lung, stomach, heart, duodenum, liver, gall bladder, intestine, kidney, adrenal gland, urinary bladder, urethra, colon, testicle, ovary, uterus, vagina, muscle, tendon, ligament, fat, cartilage, bone, endothelial cells, uterine cervix, lymph, thyroid, pituitary gland, cerebellum, and breast.

In a particular embodiment, the tissue type or cell type is blood, saliva, semen, or epidermis.

In another embodiment, the pathological condition is cancer, inflammation, autoimmune disorder, metabolic disorder, infection, degenerative disease, hormonal imbalances, a

disorder characterized by abnormal DNA methylation, neurodevelopmental disorder of ICF (immunodeficiency, centromeric instability, and facial anomalies), Rett syndrome, and fragile X syndrome.

In one embodiment, the prenatal condition in Prader-Willi syndrome, Angelman syndrome, Beckwith–Wiedemann syndrome, fragile X syndrome, Russell-Silver syndrome, Transient neonatal diabetes mellitus, Albright hereditary osteodystrophy, McCune-Albright syndrome, Familial nonchromaffin paraganglioma, Maternal and paternal UPD14 syndromes.

In another embodiment, the reference values against which the signal ratios are compared represent categories that are mixtures of other categories.

In another embodiment, the reference values against which the signal ratios are compared are from mixtures of semen and non-semen at various ratios.

In one embodiment, the methylation-sensitive restriction endonuclease is selected from the group consisting of AatII, Acc65I, AccI, AclI, ACII, AfeI, AgeI, ApaI, ApaLI, AscI, AsiSI, AvaI, AvaII, BaeI, BanI, BbeI, BceAI, BcgI, BfuCI, BglI, BmgBI, BsaAI, BsaBI, BsaHI, BsaI, BseYI, BsiEI, BsiWI, BslI, BsmAI, BsmBI, BsmFI, BspDI, BsrBI, BsrFI, BssHII, BssKI, BstAPI, BstBI, BstUI, BstZ17I, Cac8I, ClaI, DpnI, DrdI, EaeI, EagI, EagI-HF, EciI, EcoRI, EcoRI-HF, FauI, Fnu4HI, FseI, FspI, HaeII, HgaI, HhaI, HincII, HincII, HinfI, HinPII, HpaI, HpaII, Hpy166ii, Hpy188iii, Hpy99I, HpyCH4IV, KasI, MluI, MmeI, MspAII, MwoI, NaeI, NarI, NgoNIV, Nhe-HFI, NheI, NlaIV, NotI, NotI-HF, NruI, Nt.BbvCI, Nt.BsmAI, Nt.CviPII, PaeR7I, PleI, PmeI, PmlI, PshAI, PspOMI, PvuI, RsaI, RsrII, SacII, SalI, SalI-HF, Sau3AI, Sau96I, ScrFI, SfiI, SfoI, SgrAI, SmaI, SnaBI, TfiI, TscI, TseI, TspMI, and ZraI.

In a particular embodiment, the methylation-sensitive restriction endonuclease is HhaI.

In another embodiment, the methylation-dependent restriction endonuclease is selected from the group consisting of McrBC, McrA, and MrrA.

In one embodiment, the reference DNA category is natural DNA or artificially-synthesized DNA, wherein at least one genomic locus in the natural DNA or artificially-synthesized DNA comprises a core short tandem repeat (STR) locus used for DNA profiling.

In another embodiment, the genomic locus comprises a human locus selected from the group consisting of D16S539 (SEQ ID NO. 1), D7S820 (SEQ ID NO. 2), D13S317 (SEQ ID NO. 3), D5S818 (SEQ ID NO. 4), CSF1PO (SEQ ID NO. 5), TPOX (SEQ ID NO. 6), TH01 (SEQ ID NO. 7), vWA (SEQ ID NO. 8), FGA (SEQ ID NO. 9), D21S11 (SEQ ID NO. 10), D8S1179 (SEQ ID NO. 11), D18S51 (SEQ ID NO. 12), and D3S1358 (SEQ ID NO. 13), Penta D (SEQ ID NO. 14), Penta E (SEQ ID NO. 15), and Amelogenin (SEQ ID NOs. 16 and 17), D2S1338 (SEQ ID No. 18), D19S433 (SEQ ID No. 19), ACTBP2SE33 (SEQ ID No. 20), D10S1248 (SEQ ID No. 21), D1S1656 (SEQ ID No. 22), D22S1045 (SEQ ID No. 23), D2S441 (SEQ ID No. 24), and D12S391 (SEQ ID No. 25).

In one embodiment, the primers for amplification of at least one of the genomic loci are available in:

(1) a Promega Corporation commercial kit selected from the group consisting of PowerPlex® 16 HS (Cat.# DC2100, DC2101), PowerPlex® 16 (Cat.# DC6530, DC6531), PowerPlex® 2.1 (Cat.# DC6470, DC6471), PowerPlex® 16 BIO (Cat.# DC6540, DC6541), and PowerPlex® ES Systems (Cat.# DC6730, DC6731);

(2) an Applied Biosystems commercial kit selected from the group consisting of SGM, SGM+, AmpFlSTR Identifiler, AmpFlSTR Profiler, AmpFlSTR ProfilerPlus, AmpFlSTR ProfilerPlusID, AmpFlSTR SEfiler, AmpFlSTR SEfiler Plus, AmpFlSTR Cofiler, AmpFlSTR Identifiler Direct, AmpFlSTR Identifiler Plus, AmpFlSTR NGM, AmpFlSTR Y-filer, and AmpFlSTR Minifiler; or

(3) Investigator ESSplex, Investigator ESSplex SE, Investigator Nonaplex ESSplex, Investigator Hexaplex ESSplex, Investigator Triplex AFS QS, Investigator Triplex DSF, Investigator IDplex, Investigator Decaplex SE, Investigator HDplex, Investigator Argus X-12, Investigator Y-12 QS, Investigator DIPplex.

In another embodiment, one of the genomic loci is known to be unmethylated in all tissues of natural DNA. In one embodiment, one of the genomic loci is selected from the group consisting of Hypo23, Hypo28, and Hypo33.

In another embodiment, the method further comprises calculating at least one of the following ratios:

- (1) D3S1358/D18S51,
- (2) D3S1358/D7S820,
- (3) D3S1358/Penta_D,
- (4) D3S1358/TPOX,
- (5) D3S1358/FGA,
- (6) TH01/Penta_D,
- (7) D21S11/D18S51,
- (8) D21S11/D7S820,
- (9) D21S11/Penta_D,
- (10) D21S11/AMEL,
- (11) D21S11/TPOX,
- (12) D21S11/FGA,
- (13) D5S818/D18S51,
- (14) vWA/D18S51,
- (15) D5S818/Penta_E,
- (16) vWA/Penta_E,
- (17) D5S818/D7S820,
- (18) D5S818/Penta_D,
- (19) D5S818/TPOX,
- (20) D5S818/FGA,
- (21) D13S317/D7S820,

- (22) D13S317/Penta_D,
- (23) D13S317/TPOX,
- (24) D13S317/FGA,
- (25) D16S539/D7S820,
- (26) CSF1PO/D7S820,
- (27) vWA/D7S820,
- (28) D8S1179/D7S820,
- (29) D16S539/TPOX,
- (30) D16S539/FGA,
- (31) CSF1PO/Penta_D,
- (32) CSF1PO/TPOX,
- (33) vWA/Penta_D,
- (34) AMEL/TPOX,
- (35) AMEL/FGA,
- (36) vWA/D8S1179,
- (37) vWA/TPOX,
- (38) vWA/FGA,
- (39) D8S1179/TPOX, and
- (40) D8S1179/FGA.

In a particular embodiment, at least one of the genomic loci is TPOX (SEQ ID NO. 6) or FGA (SEQ ID NO. 9).

In another embodiment, the amplified loci are 90 bps or smaller in size.

In one embodiment, signal intensity is the amplification product's fluorescence level measured during capillary electrophoresis.

In one embodiment, steps (B) and (C) of the method disclosed above are performed by real-time PCR. That is, in one embodiment, all of the required reagents and primers and enzymes necessary for DNA digestion and amplification by PCR are present together in the same tube or vessel.

In one embodiment, fluorescently-labeled primers are used to PCR amplify the chosen loci of the digested DNA, wherein the signal intensity of each amplification product is the fluorescence level of each amplification product.

In another embodiment, the method further comprises hybridizing a fluorescently-labeled probe to an amplification product, wherein the signal intensity of the amplification product is the fluorescence level measured after hybridization of the probe to the product.

In another embodiment, the method further comprises assigning confidence levels to potential categories of a DNA sample, wherein the confidence level reflects the likelihood that a particular DNA category indicates the categorical source of the DNA sample; wherein the likelihood is calculated by:

(A) assigning a probability score to each comparison of a signal ratio to the reference values that correspond to the various categories of DNA, wherein the probability score is obtained by (i) assigning a gamma probability distribution function to the reference values of each potential category; wherein (ii) the probability score is equal to the value of the gamma probability distribution function assigned in step (A) at the observed signal ratio;

(B) for each category, calculating the category likelihood score, which is the product of the probability scores obtained in step (A); and

(C) normalizing likelihood scores by dividing the category likelihood score of each potential category by the sum of all category likelihood scores;

wherein the confidence level of each category is the normalized likelihood score of that category.

In another embodiment, the method further comprises simultaneously performing DNA profiling with the DNA categorization.

In one embodiment, the steps of DNA digestion and polymerase chain reaction are performed together in a single tube.

In one embodiment, the genomic loci that are amplified are chosen to produce distinct signal ratios for different potential categories.

Another aspect of the present invention is a kit for categorization of a DNA sample into at least two predetermined categories and for obtaining an associated categorization confidence level, comprising (1) primers for amplification of specific genomic loci; and at least one or more reagents selected from the group consisting of (2) a reaction buffer, (3) control DNA, (4) a methylation sensitive restriction endonuclease and/or a methylation dependent restriction endonuclease, (5) a written protocol for performing categorization. In one embodiment of the present invention, the written protocol may comprise instructions for performing any of the methods disclosed herein, including but not limited to, DNA digestion parameters, PCR cycling parameters, signal ratio analysis, and categorization analysis, as well as DNA profiling methods.

In another embodiment, the kit further comprises analysis software for performing categorization analyses.

Another aspect of the present invention is a kit for categorizing a DNA sample as semen or non-semen and for obtaining an associated categorization confidence level, comprising:

(a) primer mix, comprising the following primers:

SD1f (AAGAGCCCATCAGGCAGGTC);

SD1r (GTTTCTTGTCGAGCAGCACGTGGATGATG);

SD2f (CTCCAGAACTGGAACTTCCTG);

SD2r (GTTTCTTAACTTGGAGACGACGGCATC);

SD3f (TGGAGGACAATGCCCTGGTG);

SD3r (GTTTCTTGGCTTCACCTGCGACCGTCTC);

SD4f (CCCTCCGAGTGGCCAGCAG);

SD4r (GTTTCTGACCACTGCCGTGGGAATG);

SD5f (CTTCTCAGCCAATGGGAAGAG);

SD5r (ACGTAGAAGGACCCGAGGAC);

SD6f (TACAGACAAATCACTCAGCAGC); and

SD6r (GTTTCTTGTCTGACACTCGGTTGTAGGTATT);

(b) reaction buffer;

(c) HhaI restriction endonuclease;

(d) DNA polymerase

(d) a written protocol for performing categorization.

In another embodiment, the kit further comprises control DNA samples.

In another aspect of the present invention is a kit for categorizing a DNA sample as semen or non-semen and for obtaining an associated categorization confidence level, comprising at least one pair of forward (f) and reverse (r) primer pair combinations selected from the group consisting of:

(1) SD1f (AAGAGCCCATCAGGCAGGTC) and SD1r (GTTTCTTGTCTGAGCAGCACGTGGATGATG);

(2) SD2f (CTCCAGAACTGGAACTTCCTG) and SD2r (GTTTCTTAACTTGGAGACGACGGCATC);

(3) SD3f (TGGAGGACAATGCCCTGGTG) and SD3r (GTTTCTTGGCTTCACCTGCGACCGTCTC);

(4) SD4f (CCCTCCGAGTGGCCAGCAG) and SD4r (GTTTCTGACCACTGCCGTGGGAATG);

(5) SD5f (CTTCTCAGCCAATGGGAAGAG) and SD5r (ACGTAGAAGGACCCGAGGAC);

(6) SD6f (TACAGACAAATCACTCAGCAGC) and SD6r (GTTTCTTGTCTGACACTCGGTTGTAGGTATT).

In one embodiment, the concentration of the primers in the primer mix are: 0.6 μ M SD1f, 0.6 μ M SD1r, 1.75 μ M SD2f, 1.75 μ M SD2r, 1.25 μ M SD3f, 1.25 μ M SD3r, 1.75 μ M SD4f, 1.75 μ M SD4r, 1.75 μ M SD5f, 1.75 μ M SD5r, 0.9 μ M SD6f, and 0.9 μ M SD6r.

In one embodiment, the reaction buffer comprises 150mM TRIS-HCl, 15mM MgCl₂, 0.2mM each dntp, and 0.1 μ g/ μ l BSA.

In one embodiment, the kit further comprises a DNA ladder.

In one embodiment, the kit further comprises analysis software for performing categorization analyses.

Another aspect of the present invention is a kit for profiling, categorizing a DNA sample as semen or non-semen, and obtaining an associated categorization confidence level, comprising:

- (a) primers for amplifying at least one semen-specific locus;
- (b) primers for amplifying at least one locus used for DNA profiling;
- (c) reaction buffer;
- (d) HhaI restriction endonuclease,
- (e) DNA polymerase
- (f) a written protocol for performing categorization.

In one embodiment, the kit further comprises control DNA samples.

Another aspect of the present invention is a kit for profiling, categorizing a DNA sample as semen or non-semen, and obtaining an associated categorization confidence level, comprising:

- (a) primers for amplifying at least one semen-specific locus; and
- (b) primers for amplifying at least one locus used for DNA profiling.

In one embodiment, at least one semen-specific locus amplified by a kit disclosed herein is the L68346 locus. In one embodiment, the primers for amplifying a 70 bp semen-specific amplification product from L68346 are a forward primer comprising the sequence of CAGCAACAGCACCCAGCTTG (FAM) and a reverse primer comprising the sequence of CACAGGCTCAGTCGCGGATC.

In one embodiment, at least one semen-specific locus amplified by a kit disclosed herein is the L16264 locus. In one embodiment, the primers for amplifying a 95 bp semen-specific amplification product from L16264 are a forward primer comprising the sequence of GGACGAGTTAACTTCCTTAATTTC (FAM) and reverse primer comprising the sequence of GTTTCCTTCGCGGAACCTGGTTTAACTTC.

In another embodiment, the reaction buffer comprises 150mM TRIS-HCl, 15mM MgCl₂, 0.2mM each dntp, and 0.1µg/µl BSA.

In another embodiment, the kit further comprises at least one of (a) a DNA ladder, (b) a Material Safety Data Sheet (MSDS), and (c) analysis software for performing categorization analyses.

Another aspect of the present invention is a kit for categorizing a DNA sample as blood, saliva, semen, or skin epidermis, and for obtaining an associated categorization confidence level, comprising:

(a) primer mix that comprises forward and reverse primers for amplifying the denoted loci as follows:

1. L91762 (forward GCAGCAGGCCGCGGAGAAG (FAM); reverse AGCAGCTGTGCCGGGCCAG)

2. L68346 (forward CAGCAACAGCACCCAGCTTG (JOE); reverse CACAGGCTCAGTCGCGGATC)

3. L50468 (forward AGGAAACCTCAGTAGCAAAATTG (JOE); reverse GCGAGACTTTAGGTGTGCATC)

4. L14432 (forward CGTAGGCTGCGGTGAGCTC (FAM); reverse GATCCATGCCCCGCTGGGATG)
5. L4648 (forward CAGCCTAGACGTCAAGTTACAG (JOE); reverse ACGACCTCCGGATCCAAGT)
6. L39664 (forward CCCAGCTGGTTGGACATGTTG (FAM); reverse CACTTCCTTCGTGGACGCC)
7. L30139 (forward GAGAAGCGGGAGGATGAGAC (FAM); reverse CCGCATCTCCTCCGTCCTG)
8. L55429 (forward GCCTTCAGCAGGAAGTCCAC (JOE); reverse CCTGTGCCTCACACAGACTC)
9. L62086 (forward GTGCATGGTGTCTGGTACTTC (FAM); reverse GAAGCTCTCGTGGACTACTTG)
10. L76138 (forward CAGCCTGCTCTTCACTGCAG (JOE); reverse AGAGGCCGATGAAGCCGTAG)
11. L15952 (forward CTCCCTGATTTACGACAAGTTC (FAM); reverse GACAGTATGCTGATGCTTCTTG)
12. L36599 (forward AAGGGCAGAGTTCCGCTGTC (FAM); reverse CGGATGCAGGAGGATCCTAG)
13. L26688 (forward CGGACCAGATTGCTGGTCAC (JOE); reverse CGACCTTGCCAGATGTTTGAC)
14. L81528 (forward AGCCTCATCCACACTGACCAG (JOE); reverse TCAGAGCTCTCCTATCTGGAC)
15. L36556 (forward GCCAGGCCGTTGATGATGAC (JOE); reverse GAATATGGAGCCCTGGGCAG)

(b) reaction buffer;

(c) HhaI restriction endonuclease;

(d) DNA polymerase

(e) a written protocol for performing categorization.

In another embodiment, the kit further comprises control DNA samples.

In one embodiment, the reaction buffer comprises 150mM TRIS-HCl, 15mM MgCl₂, 0.2mM each dntp, and 0.1µg/µl BSA.

In another embodiment, the kit further comprises at least one of (a) a DNA ladder, (b) a Material Safety Data Sheet (MSDS), and (c) analysis software for performing categorization analyses.

In one embodiment in any of the methods disclosed herein, the DNA sample comprises a mixture of DNA samples.

Another aspect of the present invention is a method for calculating a distance measure of differential methylation between DNA samples, comprising:

(A) digesting each DNA sample with a methylation-sensitive and/or methylation-dependent restriction endonuclease;

(B) performing PCR on each digested DNA amplifying at least two genomic loci of which at least one is a restriction locus;

(C) for each DNA sample, determining the intensity of the signal of each amplification product;

(D) for each DNA sample, calculating “signal ratios” between the intensities of the signals produced by the loci;

(E) calculating the differential methylation measure between DNA samples by performing a quantitative comparison of their corresponding signal ratios.

In one embodiment of this method, the distance measure is sum of absolute differences between the signal ratios of the DNA samples.

In another embodiment of this method, the distance measure is the square-root of the sum of squared differences between the signal ratios of the DNA samples.

In one embodiment of this method, the distance measured between a subject sample and a reference healthy sample indicates the amount of medication required to treat a pathological state.

In another embodiment of this method, the distance measured between a DNA sample obtained from cultured cells to a reference sample indicates the number of sub-culturing procedures that the cells underwent.

Another aspect of the present invention is a primer selected from the group consisting of:

1. GCAGCAGGCCGCGGAGAAG;
2. AGCAGCTGTGCCGGGCCAG;
3. CAGCAACAGCACCCAGCTTG;
4. CACAGGCTCAGTCGCGGATC;
5. AGGAAACCTCAGTAGCAAATTG;
6. GCGAGACTTTAGGTGTGCATC;
7. CGTAGGCTGCGGTGAGCTC;
8. GATCCATGCCCGCTGGGATG;
9. CAGCCTAGACGTCAAGTTACAG;
10. ACGACCTCCGGATCCAACTG;
11. CCCAGCTGGTTGGACATGTTG;
12. CACTTCCTTCGTGGACGCC;
13. GAGAAGCGGGAGGATGAGAC;
14. CCGCATCTCCTCCGTCCTG;
15. GCCTTCAGCAGGAAGTCCAC;

16. CCTGTGCCTCACACAGACTC;
17. GTGCATGGTGTCTGGTACTTC;
18. GAAGCTCTCGTGGACTACTTG;
19. CAGCCTGCTCTTCACTGCAG;
20. AGAGGCCGATGAAGCCGTAG;
21. CTCCTGATTTACGACAAGTTC;
22. GACAGTATGCTGATGCTTCTTG;
23. AAGGGCAGAGTTCCGCTGTC;
24. CGGATGCAGGAGGATCCTAG;
25. CGGACCAGATTGCTGGTCAC;
26. CGACCTTGCCAGATGTTTGAC;
27. AGCCTCATCCACACTGACCAG;
28. TCAGAGCTCTCCTATCTGGAC;
29. GCCAGGCCGTTGATGATGAC; and
30. GAATATGGAGCCCTGGGCAG
31. AAGAGCCCATCAGGCAGGTC
32. GTTTCTTGTCGAGCAGCACGTGGATGATG
33. CTCCAGAACTGGAACCTCCTG
34. GTTTCTTAACTTGGAGACGACGGCATC
35. TGGAGGACAATGCCCTGGTG
36. GTTTCTTGGCTTCACCTGCGACCGTCTC

37. CCCTCCGAGTGGCCAGCAG
38. GTTTCTGACCACTGCCGTGGGAATG
39. CTTCTCAGCCAATGGGAAGAG
40. ACGTAGAAGGACCCGAGGAC
41. TACAGACAAATCACTCAGCAGC
42. GTTTCTTGTCTGACACTCGGTTGTAGGTATT
43. GGACGAGTTAACTTCCTTAATTTC
44. GTTTCTTCGCGGAACCTGGTTTAACTTC

Another aspect of the present invention is a primer pair for amplifying a specific human genomic locus selected from the group consisting of:

1. L91762 (forward GCAGCAGGCCGCGGAGAAG (FAM); reverse AGCAGCTGTGCCGGGCCAG);
2. L68346 (forward CAGCAACAGCACCCAGCTTG (JOE); reverse CACAGGCTCAGTCGCGGATC);
3. L50468 (forward AGGAAACCTCAGTAGCAAAATTG (JOE); reverse GCGAGACTTTAGGTGTGCATC);
4. L14432 (forward CGTAGGCTGCGGTGAGCTC (FAM); reverse GATCCATGCCCCGCTGGGATG);
5. L4648 (forward CAGCCTAGACGTCAAGTTACAG (JOE); reverse ACGACCTCCGGATCCAACCTG);
6. L39664 (forward CCCAGCTGGTTGGACATGTTG (FAM); reverse CACTTCCTTCGTGGACGCC);
7. L30139 (forward GAGAAGCGGGAGGATGAGAC (FAM); reverse CCGCATCTCCTCCGTCCTG);

8. L55429 (forward GCCTTCAGCAGGAAGTCCAC (JOE); reverse CCTGTGCCTCACACAGACTC);

9. L62086 (forward GTGCATGGTGTCTGGTACTTC (FAM); reverse GAAGCTCTCGTGGACTACTTG);

10. L76138 (forward CAGCCTGCTCTTCACTGCAG (JOE); reverse AGAGGCCGATGAAGCCGTAG);

11. L15952 (forward CTCCCTGATTTACGACAAGTTC (FAM); reverse GACAGTATGCTGATGCTTCTTG);

12. L36599 (forward AAGGGCAGAGTTCCGCTGTC (FAM); reverse CGGATGCAGGAGGATCCTAG);

13. L26688 (forward CGGACCAGATTGCTGGTCAC (JOE); reverse CGACCTTGCCAGATGTTTGAC);

14. L81528 (forward AGCCTCATCCACACTGACCAG (JOE); reverse TCAGAGCTCTCCTATCTGGAC);

15. L36556 (forward GCCAGGCCGTTGATGATGAC (JOE); reverse GAATATGGAGCCCTGGGCAG).

16. SD1 (forward AAGAGCCCATCAGGCAGGTC (FAM); reverse GTTCTTGTGCGAGCAGCACGTGGATGATG);

17. SD2 (forward CTCCAGAACTGGAACTTCCTG (FAM); reverse GTTCTTAACCTGGAGACGACGGCATC);

18. SD3 (forward TGGAGGACAATGCCCTGGTG (FAM); reverse GTTCTTGGCTTCACCTGCGACCGTCTC);

19. SD4 (forward CCCTCCGAGTGGCCAGCAG (FAM); reverse GTTCTGACCACTGCCGTGGGAATG);

20. SD5 (forward CTTCTCAGCCAATGGGAAGAG (FAM); reverse ACGTAGAAGGACCCGAGGAC);

21. SD6 (forward TACAGACAAATCACTCAGCAGC (FAM); reverse GTTTCCTGTCTGACACTCGGTTGTAGGTATT); and

22. L16264 (forward GGACGAGTTAACTTCCTTAATTTC (FAM); reverse GTTTCCTTCGCGGAACCTGGTTTAACTTC).

BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1A-D. Tissue identification assay. (A) Schematic overview of the assay. (B) Biochemical procedure – methylated loci remain intact during digestion and subsequently are amplified efficiently in the PCR, producing a strong signal (locus A) while unmethylated loci are digested and subsequently amplify inefficiently in the PCR, producing a weak signal (Locus B). The signal ratio (SR; rfu of locus A/rfu of locus B) reflects the differential methylation level between loci A and B. (C) Signal ratios between locus 1 and locus 2 are different in blood, saliva, skin, and semen, reflecting the differential methylation patterns in these tissues. (D) The observed differences in signal ratios between blood and semen (i.e. the signal) are more than a magnitude greater than the observed differences in signal ratios obtained from different PCRs and from different amounts of input DNA (i.e. the noise).

Figure 2. Tissue identification of different tissue types. Electropherograms of 8 samples (2 samples from each of 4 different tissues) are depicted demonstrating the distinct amplification pattern observed for each tissue type.

Figures 3A-D. Combined semen detection and DNA profiling. (A,B) Electropherograms of the combined assay performed on a semen (A) and urine (B) sample from the same individual. The ProfilerPlus profiles of the samples are identical, and the differences in tissue source are indicated by the differential amplification of two semen detection loci (arrows). (C) FAM channel data of the combined assay performed on samples from different tissue types. The signal ratio between the semen detection loci (arrows) is 25.04 in the semen sample, and less than 0.1 in all other tissue types. The top panel includes undigested DNA for reference. (D) FAM channel data of the combined assay performed on mixtures of semen and saliva (from different individuals) with various ratios. The signal ratio is correlated to the percentage of semen in the mixture.

Figure 4. Assessment of HhaI digestion. DNA samples of different sources were subjected to the same single digestion-amplification reaction used in the tissue identification

experiments. In this experiment, the panel of loci consisted of two loci: an undigested control locus (L98328) that lacks an HhaI site and a digested control locus (SW14) that is unmethylated in these tissues. All samples showed presence of the undigested control locus and complete absence of the digested control locus, indicating complete digestion by HhaI in contrast to the undigested sample which showed presence of both loci.

Figure 5. Aged samples. The stand-alone tissue identification assay was tested on DNA samples extracted from 20-month old stains of blood and semen. Both samples amplified well, similar to their fresh counterparts, and their source was correctly identified by the analysis software.

Figure 6. Simulation of degraded samples. Blood (top) and semen (bottom) samples were partially digested using DNaseI and analyzed with the combined semen detection and profiling assay. Although only partial profiles were obtained for both samples, the smaller tissue identification loci (the two loci in the top channel to the left of D3S1358) amplified successfully with the expected pattern, and the algorithm correctly identified the presence/absence of semen in these samples.

Figure 7. Table 1.

Figure 8. Table 2.

Figure 9. Table 3.

Figure 10. Table 4.

Figure 11. Table 5.

Figure 12. Stand alone semen detection assay. Sample plots of 6 DNA samples from different tissue types are depicted. Correct categorization (semen vs. non-semen) was achieved for all samples. For each sample the confidence level of the least likely category is shown (e.g. for a "semen" categorization, the p value of non-semen is shown)

DETAILED DESCRIPTION

1. Introduction

The present invention relates to methods and assays that enable distinguishing between sources of DNA using loci-specific primers and commercially available enzymes. An underlying aspect of an inventive assay is the comparison of signals from at least two loci amplified from a particular source of DNA, which ultimately yields a numerical ratio that indicates whether that source of the DNA sample is, in one embodiment, natural or artificial. The inventive assays also can be used to distinguish between, for example, different physiological and pathological sources of DNA and between different tissues.

The signal ratios employed by the assay are correlated to methylation levels at specific genomic loci, but do not indicate actual methylation levels at any genomic locus. Therefore the invention setup eliminates the requirement for external DNA species and/or controls, thereby both simplifying and increasing the accuracy of the assay.

The inventive DNA categorization assays described herein are therefore powerful, multiplex, accurate, and inexpensive techniques applicable in any setting that calls for the categorization of a DNA sample. Thus, the inventive assays can be used, for example, by the police in a forensics capacity; the health care industry for diagnostic and therapeutic purposes; in the insurance industry to verify claims pursuant to anti-discrimination genetic laws, such as the Genetic Information Nondiscrimination Act (H.R. 493); by prosecutors and defense counsel for evidentiary purposes in criminal trials and civil proceedings and appeals; and the food and agriculture industry to verify the integrity of meats, crops, and plants such as grapevines and sources of coffee. The present invention is not limited to this non-exclusive, but representative, list of applications.

An underlying principle of a method disclosed herein therefore is the measurement of signal intensities between amplified genomic loci and their subsequent pairwise comparison against each other to produce ratios of signal intensities between the amplified loci. For example, if loci A, B, C, and D are all amplified according to the techniques described herein, then the signal intensity for each of A, B, C, and D is measured and recorded. Then, the signal intensities of A and B, A and C, and A and D are compared respectively and signal intensity ratios calculated for each of the A/B, A/C and A/D combinations. Then, the signal intensity ratios of B/C and B/D are calculated, and so on.

The value of each of those signal intensity ratios can then be compared against the values of known reference signal intensity ratios of DNAs whose source of origin is known. Thus, for instance, in the case of DNA whose source of origin is known to be human semen, the signal intensity ratios of loci, such as the pairwise ratios of the A, B, C, and D illustrative loci are known, or can be readily established afresh when analysing a test DNA sample. Thus the DNA category is “semen” and its respective signal intensity ratios for those loci are reference values against which the test DNA sample may be compared.

Accordingly, after establishing the various signal intensity ratios of loci A, B, C, and D, from the DNA sample, each ratio can be compared against one or more reference values of particular known categories of DNA, e.g., against the semen DNA ratio values. The ratio of any loci combination from the DNA sample that more closely approximates the signal intensity ratio reference value of a particular DNA category means it is highly likely that the tested DNA sample originated from that particular category of DNA to which it more or most closely approximates when compared. Thus, a signal intensity ratio from the tested DNA sample of the A/B loci that is the same as, or close to, the value of the signal intensity ratio of the A/B loci from semen indicates that the tested DNA sample likely originated from semen. By conducting the same comparative analysis across many different loci, and against many different reference categories, the likelihood that the tested DNA sample originated from a particular biological source, or that it did not originate from a particular biological source, increases.

Similarly, in a situation where a sample might comprise mixtures of DNA from different sources, such as saliva and blood, or saliva, blood, epidermis, and semen, it will be possible, by performing the methods disclosed herein, to identify the presence of different categories of DNAs within the sample. Thus, the identification of multiple as well as sole sources of DNA present in a sample can be achieved using the methods and reagents disclosed herein.

Accordingly, the present “categorization” methods can also be thought of as a tissue identification assay. In one example, as explained in the following passage, a tissue identification assay uses a panel of loci that are differentially methylated between tissues to determine the most probable source tissue of a DNA sample. See Table 1 (Figure 7). A scheme of the assay is presented in Figure 1A. DNA from a forensic sample is digested with, for example, the HhaI methylation-sensitive restriction enzyme, which cleaves DNA at its

recognition sequence GCGC only if it is unmethylated (while leaving methylated targets intact).

A panel of tissue identification loci is then amplified by PCR from the digested DNA using fluorescently-labeled primers, and an aliquot of amplified products is separated by capillary electrophoresis. In the situation with HhaI-digested DNA, loci with higher methylation levels are amplified with higher efficiency because more DNA molecules are protected from digestion, producing a relatively strong signal in the electropherogram (Figure 1B, locus A). Conversely, loci with a lower methylation level are amplified with lower efficiency, yielding a relatively weak signal in the electropherogram (Figure 1B, locus B).

Automated signal analysis software disclosed herein enables analysis of an output electropherogram, and assigns heights (in rfu) to amplicons corresponding to tissue identification loci. Thus, the height of a single locus is correlated with its methylation level. Ratios of methylation levels between co-amplified loci can then be calculated as described generically above. For each pair of loci a signal ratio, calculated as the ratio between the heights of the first and second loci, and this ratio reflects the ratio between the methylation levels of the corresponding loci.

All calculated signal ratios can then be combined into a single numerical value. For example, 105 pairwise ratios can be calculated from a panel of 15 loci (*e.g.*, as between locus 1 and 2, locus 1 and 3, etc.). Those can then be compared to a database of reference values obtained from a data set of samples of known tissue origin. The tissue identification algorithm calculates for each potential tissue source a likelihood score, reflecting the likelihood that the DNA sample originated from that tissue, and the output of the algorithm is the most likely tissue. Disclosed elsewhere herein are algorithms and specific calculations for deriving likelihood and probability scores from these ratios and comparisons.

As shown in Example 1, for instance, such a tissue identification assay was performed on 50 samples: 14 blood, 14 saliva, 11 semen, and 11 skin epidermis using a panel of 15 tissue identification loci, ranging in amplicon size from 66 to 150 bps. Figure 2 depicts electropherograms of eight of these samples – two of each tissue type. Each tissue type had a distinct methylation profile. For example, the ratio of the L91762/L68346 loci in semen samples ranged from 0.04-0.53 and was higher in all other tissue samples (2.15-18.28; see Table 2 in Figure 8). Therefore a *low* L91762/L68346 ratio was found to be distinctive of

semen samples. The ratio of the L76138/L26688 loci, however, was low in blood and saliva (0.08-1.54) and *higher* in semen and skin epidermis (2.04-19.56; Table 2 in Figure 8). Therefore a high ratio of L91762/L68346 concomitant with a high ratio of L76138/L26688 was distinctive of skin epidermis samples

The tissue identification algorithm disclosed herein correctly identified the true tissue source of all 50 samples. Further analysis of the data using subsets of tissue identification loci revealed that the full set of 15 loci was redundant and 100% identification was achieved by using only 7 out of the 15 loci (L91762, L68346, L50468, L14432, L30139, L15952, and L26688).

Accordingly, this illustrative example corroborates the generic inventive aspect of the present invention, namely that the comparative signal intensity ratio analyses provides exacting and correct DNA sample identification and categorization.

Furthermore, another significant aspect of the present invention is that it can readily complement and expand the usefulness of existing commercial DNA profiling kits to do more than profile a particular subject's DNA. The combination of the inventive assays disclosed herein, such as the signal ratio assay described in detail below, with Promega Corporation's PowerPlex[®] 16 kit, for example, enables a user to not only profile an individual's DNA composition but also to determine whether the profiled DNA is from a natural source or has been artificially synthesized. In another example, combining DNA profiling with DNA categorization enables a user to both profile DNA and determine the source tissue of the DNA sample.

For instance, one assay described herein may employ *both* conventional profiling of a DNA sample, for example with the PowerPlex 16 kit (Promega), *and* categorization of DNA into categories of natural vs. artificial. See, for instance, Example 6 below.

For illustrative purposes only of how combined categorization and profiling may work together as disclosed herein, two samples of DNA are analyzed: (1) a natural DNA sample extracted from the blood of subject A, and (2) an artificial DNA sample synthesized by in vitro multiple displacement amplification using a commercial kit from a minute quantity of subject A's real DNA as template.

Furthermore, this illustration demonstrates how the present inventive method is far superior to any existing method because it is able to distinguish artificial DNA from natural DNA samples *even though* conventional profiling kits would identify both samples as identical to one another – because both the natural and artificial DNA are from subject A. Accordingly, notwithstanding the fact that the profiling kit would identify the samples as the same, the present inventive signal intensity ratio analyses can readily tell them apart and identify one profile as generated from artificial DNA, while the other profile was generated from natural DNA.

To explain further, the two samples are first profiled with PowerPlex16 and analyzed using a conventional profiling software, such as GeneMapperID-X. The scheme of this procedure can be briefly described as (1) performing multiplex PCR using the PowerPlex16 primer mix on a DNA sample; (2) separating the amplification products on a capillary electrophoresis machine; and then (3) analyzing the output data with, for instance, the GeneMapperID-X software. For each of the samples, the end product of this assay is a profile. The profiles of the two samples are identical. See Figures 9A-B.

Furthermore, the GeneMapperID-X software determined the profiles of both samples to be single contributor (i.e. from a single person) profiles, with no anomalies. This shows that a profile obtained from artificial DNA can be identical to a profile obtained from natural DNA. See D. Frumkin, *et al.*, *Authentication of forensic DNA samples*, Forensic Sci. Int. Genet. (2009), doi:10.1016/j.fsigen.2009.06.009, which is incorporated herein by reference.

Both samples of DNA were subsequently also analyzed by a combined profiling and categorization assay based on the following scheme: digest a DNA sample with HhaI and perform multiplex PCR using the PowerPlex16 primer mix with the addition of primers for amplifying an additional locus, such as Hypo23; separating the amplification products on a capillary electrophoresis machine; and analyzing the output data with the capillary analyzer software, such as the software algorithm disclosed herein.

For each sample, the end product of this assay is a profile and assignment of category. As can be seen in Figure 9 C and D, the natural DNA was assigned as “natural,” in contrast to the artificial sample which was assigned as “artificial.”

In this particular example, the assignment of category was performed by analysis of three parameters: (1) SR1 (signal ratio of TPOX/D8S1179), (2) SR2 (signal ratio of D3S1358/Hypo23), and (3) RR (representation ratio of OCA2/D3S1358).

The threshold values for these parameters are provided in Table 6:

Table 6

	SR1	SR2	RR
Threshold for natural	≥ 0.9	≥ 10	≥ 2
Threshold for artificial	< 0.9	< 10	< 2

The observed values for these parameters, and the assigned category are provided in Table 7 (values that are above the threshold for natural DNA are shaded).

Table 7

	SR1	SR2	RR	Category
Natural DNA sample	1.23	17.54	3.31	Natural
Artificial DNA sample	0.06	2.06	2.47	Artificial

Assignment of the category was performed according to the following rule:

If all observed parameter values are above their respective thresholds for natural DNA, then assign a “natural” category, otherwise assign an “artificial” category. Thus, the present inventive signal intensity ratio analyses can readily identify sources of DNA, such as for example from natural vs. artificial sources, and in conjunction with DNA profiling techniques provides a powerful and sophisticated method for identifying genomic materials.

Specific compositions, methods, or embodiments discussed are intended to be only illustrative of the invention disclosed by this specification. Variations on these compositions, methods, or embodiments are readily apparent to a person of skill in the art based upon the teachings of this specification and are therefore intended to be included as part of the inventions disclosed herein.

In practicing the present invention, many conventional techniques in molecular biology and recombinant DNA are used. These techniques are explained in, *e.g.*, *Current Protocols in Molecular Biology*, Vols. I-III, Ausubel, Ed. (1997); Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual*, Second Ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989); *DNA Cloning: A Practical Approach*, Vols. I and II, Glover, Ed.

(1985); *Oligonucleotide Synthesis*, Gait, Ed. (1984); *Nucleic Acid Hybridization*, Hames & Higgins, Eds. (1985); *Transcription and Translation*, Hames & Higgins, Eds. (1984); Perbal, *A Practical Guide to Molecular Cloning*; the series, *Meth. Enzymol.*, (Academic Press, Inc., 1984); *Gene Transfer Vectors for Mammalian Cells*, Miller & Calos, Eds. (Cold Spring Harbor Laboratory, NY, 1987); and *Meth. Enzymol.*, Vols. 154 and 155, Wu & Grossman, and Wu, Eds., respectively.

2. Definitions

The present technology is described herein using several definitions, as set forth throughout the specification. Unless defined otherwise, all technical and scientific terms used herein generally have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. As used herein, unless otherwise stated, the singular forms “a,” “an,” and “the” include plural reference. Thus, for example, a reference to “a nucleic acid” is a reference to one or more nucleic acids.

As used herein, the term “allele” is intended to be a genetic variation associated with a segment of DNA, *i.e.*, one of two or more alternate forms of a DNA sequence occupying the same locus.

The term “artificial DNA” or “artificially-synthesized DNA” or “artificial nucleic acid” as used herein refers to a nucleic acid which is synthesized by various *in vitro* methods. Herein disclosed are the characterizations of different methods for synthesizing DNA *in vitro*. Such *in vitro* generated nucleic acids include, but are not limited to:

1. Chemically synthesized oligonucleotides.
2. Products of PCR amplification of target sequences.
3. Products of Rolling circle amplification (RCA) of circular target sequences.
4. Products of molecular cloning (e.g. plasmids cloned in *E. coli*.)

5. DNA fragments assembled from other DNA fragments that were generated by any of methods 1-4, or a combination of them. Such assembly being achieved by any of the following methods (or a combination of them): annealing, ligation, polymerization. The

process of assembly may also include steps of breaking DNA molecules (e.g. by restriction endonucleases, mechanical shearing etc.)

6. Products of PCR-based Whole genome amplification (WGA), and/or ligation mediated PCR (LMP)-based WGA methods, including primer extension preamplification (PEP)-PCR, degenerate oligonucleotide primed (DOP)-PCR, T7-based linear amplification of DNA (TLAD), Adaptor-Ligation PCR. The Genomeplex (Sigma) commercial kit utilizes Adaptor-Ligation PCR.

7. Products of WGA by Multiple displacement amplification (MDA) and Restriction and Circularization-Aided Rolling Circle Amplification (RCA-RCA). The Repli-G (Qiagen), and GenomiPhi (GE Healthcare) commercial kits utilize this method.

8. A mix of products from any of 1-7.

9. Products from any of 1-8 in which all or some products were methylated *in vitro* following their synthesis (e.g. by SssI Methylase).

10. Products from any of 1-8 mixed with natural DNA.

11. Products from 9 mixed with natural DNA.

The term “biological sample” or “test sample” as used herein, refers to, but is not limited to, any biological sample derived from, or obtained from, a subject. The sample may comprise nucleic acids, such as DNAs or RNAs. In some embodiments, samples are not directly retrieved from the subject, but are collected from the environment, *e.g.* a crime scene or a rape victim. Examples of such samples include but are not limited to fluids, tissues, cell samples, organs, biopsies, *etc.* Suitable samples include but are not limited to are blood, plasma, saliva, urine, sperm, hair, *etc.* The biological sample can also be blood drops, dried blood stains, dried saliva stains, dried underwear stains (*e.g.* stains on underwear, pads, tampons, diapers), clothing, dental floss, ear wax, electric razor clippings, gum, hair, licked envelope, nails, paraffin embedded tissue, post mortem tissue, razors, teeth, toothbrush, toothpick, dried umbilical cord. Genomic DNA can be extracted from such samples according to methods known in the art. (for example using a protocol from Sambrook *et al.*, *Molecular Cloning: A Laboratory Manual*, Second Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989);

The term “capillary electrophoresis histogram” as used herein refers to a histogram obtained from capillary electrophoresis of PCR products wherein the products were amplified from genomic loci with fluorescent primers.

The term “methylated” as used herein means methylated at a level of at least 80% (i.e. at least 80% of the DNA molecules methylated) in DNA of cells of tissues including blood, saliva, semen, epidermis, nasal discharge, buccal cells, hair, nail clippings, menstrual excretion, vaginal cells, urine, and feces.

The term “partially-methylated” as used herein means methylated at a level between 20-80% (i.e. between 20-80% of the DNA molecules methylated) in DNA of cells of tissues including blood, saliva, semen, epidermis, nasal discharge, buccal cells, hair, nail clippings, menstrual excretion, vaginal cells, urine, and feces.

The term “unmethylated” as used herein means methylated at a level less than 20% (i.e. less than 20% of the DNA molecules methylated) in DNA of cells of tissues including blood, saliva, semen, epidermis, nasal discharge, buccal cells, hair, nail clippings, menstrual excretion, vaginal cells, urine, bone, and feces. The methods provided herein have been demonstrated to distinguish methylated and unmethylated forms of nucleic acid loci in various tissues and cell types including blood, saliva, semen, epidermis, nasal discharge, buccal cells, hair, nail clippings, menstrual excretion, vaginal cells, urine, bone, and feces.

The terms “determining,” “measuring,” “assessing,” “assaying”, and “evaluating” are used interchangeably to refer to any form of quantitative or qualitative measurement, and include determining if a characteristic, trait, or feature is present or not. Assessing may be relative or absolute. “Assessing the presence of” includes determining the amount of something present, as well as determining whether it is present or absent.

The term “forensics” or “forensic science” as used herein refers to the application of a broad spectrum of methods aimed to answer questions of identity being of interest to the legal system. For example, the identification of potential suspects whose DNA may match evidence left at crime scenes, the exoneration of persons wrongly accused of crimes, identification of crime and catastrophe victims, or establishment of paternity and other family relationships.

The term “locus” (plural – loci) refers to a position on a chromosome of a gene or other genetic element. Locus may also mean the DNA at that position. A variant of the DNA sequence at a given locus is called an allele. Alleles of a locus are located at identical sites on homologous chromosomes. A restriction locus is a locus that contains the recognition sequence of the restriction enzyme used in the assay. The term digested control locus as used herein refers to a genomic locus known to be unmethylated in all potential DNA categories. The term undigested control locus as used herein refers to a genomic locus that is either known to be methylated in all potential DNA categories, or is a genomic locus lacking the recognition sequence of the endonuclease used in the assay.

The term “natural DNA” or “natural nucleic acid” as used herein refers to, but is not limited to, nucleic acid which originates directly from the cells of a subject without modification or amplification.

The term “nucleic acid” as used herein refers to, but is not limited to, genomic DNA, cDNA, hnRNA, mRNA, rRNA, tRNA, fragmented nucleic acid, and nucleic acid obtained from subcellular organelles such as mitochondria. In addition, nucleic acids include, but are not limited to, synthetic nucleic acids or *in vitro* transcription products.

The term “nucleic-acid based analysis procedures” as used herein refers to any identification procedure which is based on the analysis of nucleic acids, *e.g.* DNA profiling.

The term “polymerase chain reaction (PCR) stutter” as used herein refers to PCR byproducts, obtained along with the main PCR product. These “stutter” byproducts are usually shorter by multiples of the repeated unit produced in the course of PCR amplification of STR sequences. The mechanism by which these artifacts are formed is understood, but it represents an intrinsic limitation of the PCR technology and therefore no effective remedy has been found to eliminate these spurious products (Olejniczak M, Krzyzosiak WJ., *Electrophoresis*. 2006 Oct; 27(19):3724-34). The term “-1 stutter” as used herein refers to a stutter byproduct that is one repeat unit smaller than its associated allele. Similarly, “+1 stutter” refers to a stutter byproduct that is one repeat unit larger than its associated allele. The term ‘-1 stutter fraction’ refers to the height (or area) of the -1 stutter peak divided by the height (or area) of the true allele peak. Similarly, “+1 stutter fraction” refers to the height (or area) of the +1 stutter peak divided by the height (or area) of the true allele peak.

The term “Restriction and Circularization-Aided Rolling Circle Amplification (RCA-RCA)” refers to a whole genome amplification procedure which retains the allelic differences among degraded amplified genomes while achieving almost complete genome coverage. RCA-RCA utilizes restriction digestion and whole genome circularization to generate genomic sequences amenable to rolling circle amplification.

The term “STR primers” as used herein refers to any commercially available or made-in-the-lab nucleotide primers that can be used to amplify a target nucleic acid sequence from a biological sample by PCR. There are approximately 1.5 million non-CODIS STR loci. Non-limiting examples of the above are presented in the following website www.cstl.nist.gov/biotech/strbase/str_ref.htm that currently contains 3156 references for STRs employed in science, forensics and beyond. In addition to published primer sequences, STR primers may be obtained from commercial kits for amplification of hundreds of STR loci (for example, ABI Prism Linkage Mapping Set-MD10 -Applied Biosystems), and for amplification of thousands of SNP loci (for example, Illumina BeadArray linkage mapping panel). The term “CODIS STR primers” as used herein refers to STR primers that are designed to amplify any of the thirteen core STR loci designated by the FBI's “Combined DNA Index System”, specifically, the repeated sequences of TH01, TPOX, CSF1PO, VWA, FGA, D3S1358, D5S818, D7S820, D13S317, D16S539, D8S1179, D18S51, and D21S11, and the Amelogenin locus.

The term “signal ratio” as used herein refers to the ratio between the intensities of the signals obtained from amplification of two loci.

3. Methods for distinguishing between natural and artificial DNA samples

In one aspect, the present invention provides a method for distinguishing between natural and artificial DNA samples. A general scheme of the invention is as follows: the method accepts as input a DNA sample. The DNA undergoes a procedure including one or more biochemical steps followed by signal detection. In the last step of the procedure, the signal is analyzed to determine whether the DNA is natural or artificial. In another aspect, the present invention provides a method for verifying that a DNA profile is of natural DNA. A general scheme of the invention is as follows: the method accepts as an input a DNA sample that underwent profiling (*e.g.* with Identifiler). The DNA sample undergoes a verification procedure which includes one or more biochemical steps followed by signal

detection. In the last step of the entire procedure, data from both profiling and verification of the DNA sample are analyzed. The signal analysis determines whether the profile obtained from the DNA sample represents natural (*in vivo*) or artificial (*in vitro*) DNA.

In various aspects, the methods of the present invention concern the verification that DNA profiles represent natural DNA. The methods are employed on a DNA sample in question, for example, DNA from a blood sample found at a crime scene. The isolation of nucleic acids (*e.g.* DNA) from a biological sample may be achieved by various methods known in the art (*e.g.* see Sambrook et al, (1989) *Molecular Cloning: A Laboratory Manual*, 2nd ed. Cold Spring Harbor, New York). Distinguishing between natural and artificial DNA, or the determination whether a DNA profile represents natural DNA, may be accomplished using various strategies, including those described in the following sections.

4. Methylation

Methylation in the human genome occurs in the form of 5-methyl cytosine and is confined to cytosine residues that are part of the sequence CG (cytosine residues that are part of other sequences are not methylated).

Some CG dinucleotides in the human genome are methylated, and others are not. In addition, methylation is cell and tissue specific, such that a specific CG dinucleotide can be methylated in a certain cell and at the same time unmethylated in a different cell, or methylated in a certain tissue and at the same time unmethylated in different tissues. Since methylation at a specific locus can vary from cell to cell, when analyzing the methylation status of DNA extracted from a plurality of cells (*e.g.* from a forensic sample), the signal can be mixed, showing both the methylated and unmethylated signals in varying ratios. The inventors discovered that in some genomic regions all CG loci are constitutively methylated. These regions are provided in Table 1 and in the section herein entitled Sequences. The inventors also discovered that in some genomic regions all CG loci are partially methylated. These regions are provided in Table 1 and in the section herein entitled Sequences. The inventors also discovered that in some genomic regions all CG loci are constitutively unmethylated. These regions are provided in Table 1 and in the section herein entitled Sequences. The inventors also discovered contiguous genomic regions containing constitutively methylated, partially methylated, and constitutively unmethylated CG loci. These regions are provided in Table 1 and in the section herein entitled Sequences. There are

several different methods for determining the methylation level of genomic loci. Examples of methods that are commonly used are bisulfite sequencing, methylation-specific PCR, and methylation-sensitive endonuclease digestion. Further, various data sources are available for retrieving or storing DNA methylation data and making these data readily available to the public, for example MetDB (<http://www.methdb.net>).

Exemplary methods for determining the methylation level of nucleic acids include, but are not limited to the following methods:

Bisulfite sequencing. Bisulfite sequencing is the sequencing of bisulfite treated-DNA to determine its pattern of methylation. The method is based on the fact that treatment of DNA with sodium bisulfite results in conversion of non-methylated cytosine residues to uracil, while leaving the methylated cytosine residues unaffected. Following conversion by sodium bisulfite, specific regions of the DNA are amplified by PCR, and the PCR products are sequenced. Since in the polymerase chain reaction uracil residues are amplified as if they were thymine residues, unmethylated cytosine residues in the original DNA appear as thymine residues in the sequenced PCR product, whereas methylated cytosine residues in the original DNA appear as cytosine residues in the sequenced PCR product.

5. Methylation-specific PCR

Methylation specific PCR is a method of methylation analysis that, like bisulfite sequencing, is also performed on bisulfite-treated DNA, but avoids the need to sequence the genomic region of interest. Instead, the selected region in the bisulfite-treated DNA is amplified by PCR using two sets of primers that are designed to anneal to the same genomic targets. The primer pairs are designed to be “methylated-specific” by including sequences complementing only unconverted 5-methylcytosines, or conversely “unmethylated-specific”, complementing thymines converted from unmethylated cytosines. Methylation is determined by the relative efficiency of the different primer pairs in achieving amplification.

It should be understood in the context of the present invention that methylation specific PCR determines the methylation level of CG dinucleotides in the primer sequences only, and not in the entire genomic region that is amplified by PCR. Therefore, CG dinucleotides that are found in the amplified sequence but are not in the primer sequences are not part of the CG locus.

6. Methylation-sensitive endonuclease digestion

Digestion of DNA with methylation-sensitive endonucleases represents a method for methylation analysis that can be applied directly to genomic DNA without the need to perform bisulfite conversion. The method is based on the fact that methylation-sensitive endonucleases digest only unmethylated DNA, while leaving methylated DNA intact. Following digestion, the DNA can be analyzed for methylation level by a variety of methods, including gel electrophoresis, and PCR amplification of specific loci.

In the procedure based on methylation-sensitive endonuclease digestion, each CG locus is comprised of one or more CG dinucleotides that are part of recognition sequence(s) of the methylation-sensitive restriction endonuclease(s) that are used in the procedure. CG dinucleotides that are found in the amplified genomic region, but are not in the recognition sequence(s) of the endonuclease(s) are not part of the CG locus.

In one embodiment, the one or more CG loci that are detected are partially methylated in natural DNA, but would be unmethylated in artificial DNA. Partial methylation would be expected to result in a mixture of T and C at the position being interrogated. Hybridization would be observed to both the T specific probes/primers and the C specific probes/primers, similar to detection of a heterozygous SNP. Relative amounts of hybridization may be used to determine the relative amount of methylation. Alternatively, both C and T would be observed upon bisulfite sequencing. Alternatively, fluorescent signals corresponding to amplification products of methylated or partially methylated CG loci can be detected.

7. Overview of the Present Signal ratio Assay

As mentioned above, one particular assay of the present invention involves the quantitative comparison of intensity of the signals from a pair of locus-specific amplification products produced by performing a Polymerase Chain Reaction on restriction-digested DNA. See Figure 1A,B. The numerical ratio of intensities corresponds to the category of the template DNA, for example whether the DNA sample was extracted from semen or from a non-semen tissue. For example, in one embodiment, locus 1 and locus 2 can be amplified using fluorescently labeled primers, separated by electrophoresis, and the intensity of the signals is the relative fluorescence units (rfu) of peaks corresponding to the loci. See Figure 1B,C. The intensity of the signals will correspond to the successfulness of amplification of locus 1 and locus 2 from the source DNA template. By comparing rfu between the two

amplification products one can calculate a signal ratio - a ratio that reflects the ratio of the quantities of the amplification products.

In addition, however, one aspect of this assay includes the predetermination of the expected signal ratios from various types of DNA categories. Thus, the template DNA that is subject to analysis is first digested with a methylation-sensitive restriction endonuclease before it is cycled through the quantitative-styled PCR amplification protocol. If the level of methylation at the site/s of the restriction recognition sequence/s in the pair of the two loci in the various categories of DNA is known, the skilled person can predict the expected signal ratios of amplification products of the pair of loci in the respective categories. Alternatively, if the levels of methylation are unknown, the skilled person can nonetheless empirically determine the signal ratios between a pair of genomic loci in the various categories of DNA by performing the assay on samples from known categories.

With this premise, the present inventive assays comprise digesting a DNA sample with a methylation-sensitive and/or methylation dependent enzyme, performing a PCR amplification reaction on the digested DNA, and determining the intensity of the signals from locus-specific amplification products. As mentioned, the intensity of signals can be quantified or measured by using fluorescent PCR. Determination of the category of the DNA sample is performed by comparing the observed signal ratios to reference ratios obtained from DNA samples of known categories, and determining which category corresponds best to the observed ratios.

This particular signal ratio assay is not reliant upon identifying or obtaining measurements of the absolute methylation fraction or level of selected loci. In addition, this particular signal ratio assay is not reliant upon the efficiencies of the primer pairs used, does not necessitate that both primer pairs have similar efficiencies, is not reliant upon amount of input template DNA, is not reliant upon specific thermocycler machine and reaction conditions. Rather, the assay determines the ratio between two signals which correspond to the ratio of methylation levels in the different loci. By this manner, the quantity or concentration of starting DNA material in the sample is irrelevant to the analysis and does not skew the output results. That is, the ratio of signal levels between a first locus and a second locus will remain constant regardless of how much DNA is used as a template for PCR and regardless of the number of amplification cycles that are run on the thermocycler. For example, a signal ratio of 10 between loci 1 and 2 will remain the same whether the input

DNA represents methylation levels of 0.9 and 0.09 (90% methylation in locus 1 and 9% in 2), or 0.5 and 0.05 (50% methylation in locus 1 and 5% in 2), etc. In another example, a signal ratio of 10 between loci 1 and 2 will remain the same regardless of the amount of DNA template used in the reaction.

The fact that the inventive assay does not depend on determination of methylation levels at any genomic locus, and does not require any external control, allows the assay to be more accurate, cost effective, and applicable to samples with low amounts of DNA.

In the context of forensic tissue identification, the inventive assay represents a standardized procedure which is identical regardless of the specific tissue identification task that is being performed (except for the use of panel-specific primers), that requires only standard forensic procedures and equipment, thus enabling the assay to be automated and easily integrated into forensic labs.

Furthermore, the setup of the inventive assay makes possible combining tissue identification and DNA profiling into a single assay.

The fact that the inventive assay performs digestion and amplification in the same reaction minimizes the complexity (the number of required tubes and pipetting steps of the assay), thereby minimizing the risk of contamination and/or error.

This inventive assay therefore provides a useful biochemical marker in the form of, in one example, a numerical ratio, that can be used to differentiate between natural DNA and artificial DNA that was synthesized *in vitro* such as by PCR, chemical synthesis, or multiple displacement amplification (MDA). More particular details of this exemplary assay follow.

8. Characteristics of Primers for Locus-Specific Amplification

Accordingly, an aspect of the present invention concerns obtaining a “signal ratio” in which the fluorescence intensities of multiple DNA loci amplification products produced from fluorescent PCR are compared to one another in order to calculate ratios between successive pairs of loci, e.g., Loci #1 vs. Loci #2; Loci #1 vs. Loci #3; Loci #1 vs. Loci #4; Loci #2 vs. Loci #3, Loci #2 vs. Loci #4, and so on. When this technique is used to test the natural vs. artificial nature of a DNA sample, the primers that are used in the signal ratio amplification reactions are chosen so as to amplify a pair of loci that are differentially

methyated in natural DNA (e.g. the first is 90% methyated and the second is 30% methyated in natural DNA).

One consideration for selecting which two pairs of primers (a first pair and a second pair) to use to amplify two loci (1) and (2) is the degree to which the two loci are differentially methyated in *in vivo* DNA. Thus, a pair of loci in which locus (1) is more methyated in comparison to locus (2), are examples of loci that can be used to design primers for the signal ratio amplification assay.

9. Selection of Genomic Loci

1. Selection of putative loci can be performed in any of three modes:

(i) Select a locus closest to the transcription start site of a gene that has been shown in the literature to be differentially expressed (mRNA data) in the relevant categories (e.g. different tissues).

(ii) Select a locus that has been shown in the literature to be differentially methyated between different categories (e.g. tissues).

(iii) Select a random genomic locus.

2. Design primers for putative loci.

3. Empirically test "informativeness" of putative loci. For each putative locus calculate the signal ratio between the putative locus and another putative locus, or alternatively, between the putative locus and an undigested control locus.

4. Choose "informative" loci (i.e. loci whose signal ratios are significantly different in the different DNA categories) for use in the assay.

10. Methylation-Sensitive Restriction Endonucleases

A second consideration of the present inventive method is the selection of loci that are or are not cut or digested by a methylation-sensitive and/or methylation-dependent restriction endonuclease, depending on their methylation status. The endonuclease is selected if, for instance, it is unable to cut the DNA strand if its recognition sequence in that locus is methyated. Thus, in the context of locus (1), which is methyated, and locus (2), which is

not methylated, an endonuclease like HhaI or HpaII will not digest locus (1) but will digest locus (2). Accordingly, the selection of loci for amplification in the signal ratio assay may also take into account the presence of methylation-sensitive restriction endonuclease recognition sequences within each locus.

Forward and reverse primers can then be designed to anneal to a region of DNA that flanks the recognition sequence of the loci.

Accordingly, in the case of a methylation-sensitive enzyme, if a locus is methylated it will (A) not be digested but (B) it will be amplified. Conversely, if a locus is unmethylated, it will (A) be digested but (B) not amplified. In the case of a methylation-dependent enzyme, the situation is vice versa.

11. Advantages of the Signal Ratio Assay

The signal ratio assay of the present invention has several advantages and benefits over other approaches for analyzing methylation, and therefore the skilled person would appreciate the various advantageous uses to which this signal ratio assay can be put. For instance, this assay is insensitive to various “noise” factors inherent when relying on the absolute intensity level of an amplicon signal that fluctuates as a consequence of changes in template DNA concentration, thermocycler manufacturer, and PCR conditions, and presence of inhibitors. Instead, the presently-calculated methylation ratio does not depend on such factors, since the two analyzed loci are co-amplified in the same reaction and are therefore subject to the effects of such disparities. Thus, the present inventive method does not require absolute quantification of genomic targets or amplicons; and the assay is a single stand-alone reaction with no need for a standard curve or any external controls.

The signal ratio can be performed on very small quantities of DNA in a single biochemical reaction and is therefore an inexpensive, rapid, and powerful method for establishing, for example, whether a sample contains natural or artificial DNA. An important feature of the design of the inventive methods is that it can be combined with other PCR-based procedures, such as DNA profiling, in a single biochemical reaction.

12. Capillary Electrophoresis and Multiplexing

The rapidity of the analysis is evident in consideration of the use of, for instance, capillary electrophoresis to separate numerous amplification products produced from the amplification of multiple pairs of target loci. As described above the present signal ratio assay can be performed on multiple loci, and in each case a signal ratio is calculated for each pair of loci separately. For example, if four loci (A,B,C,D) are co-amplified in the reaction, six different signal ratios can be calculated, *i.e.*: A/B, A/C, A/D, B/C, B/D, C/D.

Accordingly, if “n” loci are co-amplified, then $(n^2-n)/2$ different ratios can be calculated. Therefore, the amount of information that is provided by the present methylation assay rises exponentially with the number of analyzed loci. Capillary electrophoresis, as opposed to real-time PCR amplification methods, can distinguish between a large number of loci in a single run. For example, for DNA profiling, 17 genomic loci are routinely co-amplified from a particular DNA sample, and analyzed together. As a consequence, the performance of the present signal ratio assay on all 17 loci yields 136 independent signal ratios. Real-time PCR cannot simultaneously distinguish in a single reaction those numbers of discrete amplification products necessary to produce 136 ratios. About four loci can be distinguished by real time PCR, which corresponds to the calculation of only six ratios.

By contrast, capillary electrophoresis can readily separate out amplification products from all paired permutations of 17 loci and can therefore readily produce data to simultaneously calculate all 136 signal ratios in a single reaction. Theoretically, hundreds of loci can be run together and separated in a single capillary electrophoresis run. In order to facilitate analysis of multiple ratios, it can be advantageous to calculate a single score, herein termed “Combined Signal Score” (CSS), from these ratios. The CSS can then be used, for example, in order to ascertain whether the source DNA sample is natural or artificial.

13. Sequences

The sequences provided herein for the various CODIS, PowerPlex® 16, and other loci commonly used for profiling, *i.e.*, SEQ ID NOs. 1-25 are of genomic loci commonly used for STR profiling, have been analyzed herein to determine (1) the position of every cytosine-guanine (CG) dinucleotide, (2) the methylation level of all CG dinucleotides in the locus in natural DNA, and (3) the methylation sensitive and methylation dependent restriction enzyme profile for that particular locus. The sequence listing included within the text of this

application therefore provides guidance to the skilled person in the identification of particular methylation-sensitive and methylation-dependent restriction endonucleases that can be used in accordance with the inventive ratio-generating assay methods. For example, SEQ ID NO. 8, which is the 751-long nucleotide sequence for the “vWA” CODIS and PowerPlex® 16-amplified locus, comprises only one CG dinucleotide at position 571. That CG and the nucleotides that immediately flank it, provide the recognition sequence for enzymes BssKI, HpaII, Nt.CviPII, and SEQ ID NOs 26-32 are genomic loci that have been found by the inventors to produce signal ratios with significantly different values between natural and artificial DNA samples. SEQ ID NOs 33-54 are genomic loci that have been found by the inventors to produce signal ratios with significantly different values (based on the Kolomogorov-Smirnov test using a threshold of $p = 0.05$) between samples of DNA extracted from different tissues.

The sequence information provided herein also permits the skilled artisan to design forward and reverse amplification primers that anneal to regions of a selected locus that flank the CG and restriction site. Thus, the present invention is not limited to the amplification of, for instance, CODIS loci, using only those commercially available primers, although the use and availability of commercially available primers can be a more convenient and cost-effective option for performing the present inventive authentication assays.

14. Algorithm and Software

In one embodiment, calculation of signal ratios and/or representation ratios (for example in order to determine whether a DNA profile represents natural DNA) is performed based on analysis of the signal intensities of amplification products of fluorescent PCR that are run on a capillary electrophoresis machine. The output of the capillary electrophoresis machine is a binary computer file (for example, an FSA file in the case of capillary electrophoresis machines of Applied Biosystems). This file includes information regarding the capillary electrophoresis run, including the channel data, which is the relative fluorescent units (rfus) of each fluorophore as a function of each sampling time point (called datapoint).

The present invention also describes a software program that accepts as input a file that is the output a capillary electrophoresis machine run, and calculates the fluorescence intensities of a set of loci whose amplification products were run on the capillary electrophoresis machine. Based on these intensities, the software calculates signal ratios

and/or representation ratios (or combined scores based on these ratios), based on a set of predefined pairs of loci for which the ratios are defined to be calculated. Finally, the software outputs a decision based on the comparison of the calculated ratios to predefined thresholds.

Following is a scheme of this analysis performed by the software program:

1. Subtract from the signal of each dye channel the baseline level, defined as the mean rfu level of that channel in the range 400-500 bps, and remove spectral overlap between fluorophores based on the matrix obtained from the spectral calibration procedure of the capillary electrophoresis machine.
2. Obtain the rfu level of each analyzed locus, defined as the maximal rfu level in the range -0.5bp to +0.5bp of the expected size of the locus.
3. Calculate signal ratios between a set of pairs of loci, defined as the ratio between the rfu levels of the first and second loci in the pair (to avoid division by zero assign a level of 50 rfus to all loci with rfu levels below 50).
4. For each ratio calculated in step 3 and for each tissue type calculate a Probability Score, defined as the value of the reference probability function (see below) at the observed ratio.
5. For each tissue type, calculate a Combined Probability Score, defined as the product of the separate Probability Scores for that tissue obtained in step 4 (where n is the number of signal ratios).
6. For each tissue type, calculate a Likelihood Score (representing the likelihood that the DNA sample originated from that tissue), defined as the ratio between the Combined Probability Score of the tissue and the sum of the Combined Probability Scores of all tissues. The tissue with the highest Likelihood Score is considered the source tissue of the sample.

15. Reference probability functions

The reference probability function of a certain pair of loci in a certain tissue is the gamma distribution function fitted to the set of signal ratios observed in samples of natural DNA of that tissue type.

16. Correction for incomplete digestion

Since incomplete digestion of template DNA might occur as a result of the presence of inhibitors in the sample, it is useful to incorporate into the assay an analysis of actual digestion level. This can be achieved by incorporation of at least one "digested control locus" and at least one "undigested control locus". A digested control locus is a genomic locus known to be unmethylated in natural DNA in all potential tissues. An undigested control locus is a locus that is either known to be methylated in natural DNA in all potential tissues, or is a locus lacking the recognition sequence of the endonuclease used in the assay. The signal ratio between a digested control locus and an undigested control locus is correlated to the actual digestion level in the assay. The algorithm can make use of such loci by checking whether adequate digestion occurred (and if not – aborting analysis), and by modifying actual rfu levels of loci to values representing rfu values that would have been obtained had complete digestion occurred. If this modification is performed, then signal ratios are calculated from the modified rfu values.

EXAMPLES

MATERIALS AND METHODS

A. Collection of biological tissues

Blood (venous/menstrual), saliva, semen, skin epidermis, urine, and vaginal secretion were collected from volunteers. Informed consent was obtained from all participants recruited into the study.

B. DNA extraction and quantification

DNA was extracted from all samples using by organic extraction. The quantity of recovered DNA was determined using the Quantifiler[®] Human DNA quantification kit (Applied Biosystems) and the 7300 Real-Time PCR system (Applied Biosystems).

C. Selection of genetic loci and primer design

Random CpG islands (defined as a region ≥ 200 bp and with a GC content ≥ 0.5 and with at least 8 CGs; corresponding to observed/expected CpG ratio > 0.6) were searched using a software program developed for this task. From these, CpG islands that contained a HhaI recognition sequence were selected for initial screening. Primers (18-28 bps) flanking the HhaI recognition sequence were designed with a T_m of 64-66°C and for amplicon sizes of 66-150 bps. Fluorescently-labeled forward primers (FAM or JOE) were purchased from Integrated DNA Technologies. Loci were screened in pairs for differential methylation by comparing signals obtained from amplification of pooled DNA samples (each containing DNA from 10 individuals) of each tissue type. Loci that showed significant differential amplification patterns (defined as the largest signal ratio observed in all tissue types being greater than three times the smallest ratio observed) were selected for multiplexing. In total, 205 genomic loci were screened of which 38 showed significant differential amplification patterns. From these, 16 loci were used in the experiments. For the stand-alone tissue identification assay, 15 loci that could be co-amplified without significant noise were used. For the combined semen detection and profiling assay two loci (L68346 which was also used for the stand-alone assay and L16264) were used.

D. Endonuclease digestion, PCR, and capillary electrophoresis

In stand-alone assay experiments, each DNA sample was digested by HhaI and subjected to PCR amplification in the same reaction, consisting of 30 U HhaI (New England Biolabs), 2.5 U AmpliTaq Gold (Applied Biosystems), 2.5 µg BSA (New England Biolabs), 0.2 mM each dNTP, 0.1-0.3 µM each primer, 2.5µl reaction buffer (150mM Tris-Hcl, 15mM MgCl₂) and DDW to a total volume of 25µl. The reaction composition in the combined tissue identification-profiling experiments was the same as in the stand-alone assay except for the primers, which were 0.2µM each semen identification primer together with 5µl Primer Set of ProfilerPlus (Applied Biosystems). Reactions (for both stand-alone and combined assay) were performed in a GeneAmp[®] PCR System 9700 (Applied Biosystems), and the thermocycling program used was: 37°C for 15 min, 95°C for 11 min, followed by 28 cycles of 94°C for 1 min, 59°C for 1 min, 72°C for 1 min, and followed by a final extension step of 60°C for 45 min. A mix containing 1.5µl of amplification products, 24.5µl HiDi formamide (Applied Biosystems) and 0.5µl ROX size standard (Applied Biosystems) was denatured (95°C for 3 min immediately followed by 3 min on ice), and run on an ABI 310 Genetic Analyzer (Applied Biosystems) according to the manufacturer's instructions. All resulting electropherograms were analyzed using in-house developed software, and electropherograms resulting from combined tissue identification-profiling experiments were also analyzed using GeneMapper ID-X analysis software (Applied Biosystems).

E. Tissue identification assay –data analysis

Data analysis for tissue identification was performed by an in-house developed software that accepts as input fsa files and outputs the most likely source tissue. Derivation of semen percentage in mixture samples was performed based on profiling loci and tissue identification loci. Derivation based on profiling loci was performed by dividing the sum of rfu values of alleles corresponding to the contributor of the semen by the total sum of rfu values of all alleles in STR loci that had 4 genotyped alleles. Derivation based on tissue identification loci was performed first by calculating the expected signal ratio for all possible percentages of mixtures (from 0 to 100% with 1% increments, based on the signal ratio of undigested DNA), and the algorithm provided the percentage of semen for which the expected signal ratio is closest to the observed signal ratio.

F. Simulation of degraded samples

Twenty nanograms of each sample were digested for 10min at 37°C with 0.01U DNaseI (Ambion), 2µl 10X DNaseI buffer (Ambion) in total reaction volume of 20µl, followed by heat-inactivation at 75°C for 10 min. One nanogram of each digested DNA sample was then subjected to the combined tissue identification-profiling assay.

Example 1

Categorization of DNA according to tissue type – stand alone assay

In this example, the categorization procedure is termed “tissue identification assay.” The tissue identification assay uses a panel of loci that are differentially methylated between tissues to determine the most probable source tissue of a DNA sample as shown in Table 1 (Figure 7). A scheme of the assay is presented in Figure 1A. One nanogram of DNA from a forensic sample in question is digested with the HhaI methylation-sensitive restriction enzyme, which cleaves DNA at its recognition sequence GCGC only if it is unmethylated (while leaving methylated targets intact). A panel of tissue identification loci is then amplified by PCR from the digested DNA using fluorescently-labeled primers, and an aliquot of amplified products is separated by capillary electrophoresis. Loci with higher methylation levels are amplified with higher efficiency (because more DNA molecules are protected from digestion), producing a relatively strong signal in the electropherogram (Figure 1B, locus A). Conversely, loci with a lower methylation level are amplified with lower efficiency, yielding a relatively weak signal in the electropherogram (Figure 1B, locus B).

Automated signal analysis software enables analysis of an output electropherogram, and assigns heights (in rfu) to amplicons corresponding to tissue identification loci. Although the height of a single locus is correlated with its methylation level, the actual methylation level cannot be derived, because that level also depends on the precise template concentration and specific PCR conditions. Instead, ratios of methylation levels between co-amplified loci can be calculated, since all the loci are subjected to the same template concentration and specific reaction conditions. Therefore, for each pair of loci a signal ratio, defined as the ratio between the heights of the first and second loci, is calculated, and this ratio reflects the ratio between the methylation levels of the corresponding loci. All calculated signal ratios are combined into a single vector (if, for example, in a 15-loci panel the vector contains 105 ratios – between locus 1 and 2, locus 1 and 3, etc.), which is then compared to a database of

reference vectors obtained from a data set of samples of known tissue origin. The tissue identification algorithm calculates for each potential tissue source a likelihood score, reflecting the likelihood that the DNA sample originated from that tissue, and the output of the algorithm is the most likely tissue.

A stand-alone tissue identification assay was performed on 50 samples: 14 blood, 14 saliva, 11 semen, and 11 skin epidermis using a panel of 15 tissue identification loci, ranging in amplicon size from 66 to 150 bps. Figure 2 depicts electropherograms of eight of these samples – two of each tissue type. Although samples of the same tissue type from different individuals were not completely identical (due to natural variability of methylation levels and stochastic PCR effects), it was nevertheless evident that each tissue type had a distinct methylation profile. For example, the ratio of L91762/L68346 in semen samples ranged from 0.04-0.53 and was higher in all other tissue samples (2.15-18.28; Table 2 in Figure 8), and therefore a low L91762/L68346 ratio was distinctive of semen samples. The ratio of L76138/L26688 was low in blood and saliva (0.08-1.54) and higher in semen and skin epidermis (2.04-19.56; Table 2 in Figure 8); therefore a high ratio of L91762/L68346 concomitant with a high ratio of L76138/L26688 was distinctive of skin epidermis samples. The tissue identification algorithm correctly identified the true tissue source of all 50 samples. Further analysis of the data using subsets of tissue identification loci revealed that the full set of 15 loci was redundant and 100% identification was achieved by using only 7 out of the 15 loci (L91762, L68346, L50468, L14432, L30139, L15952, and L26688). Because tissue identification relies on efficient digestion, complete digestion of template DNA by HhaI was assessed in a single digestion-amplification reaction setup. The 50 DNA samples were subjected to the same procedure but using a panel of two different loci as shown in Table 3 (Figure 9) – a control locus without a HhaI site (L98328) and a locus which was previously shown to be completely unmethylated in these tissues (SW14; see D. Frumkin et al., Authentication of forensic DNA samples, *Forensic Sci. Int. Genet.* 2009, doi: 10.1016/j.fsigen.2009.06.009, which is incorporated herein by reference). All samples showed presence of the reference locus and complete absence of SW14, indicating complete digestion by HhaI in contrast to the undigested sample which showed presence of both loci (example runs are shown in Figure 4).

Example 2

Analysis of noise factors affecting signal ratios

The assay described in example 1 is based on utilizing loci that are differentially methylated between tissues, for which signal ratio values are generally tissue-specific (Figure 1C). However, even in differentially methylated tissues, there potentially could be some overlap in the observed signal ratios as a result of natural variability in methylation levels among different individuals and/or artifacts associated with the PCR such as differences in DNA template concentration (e.g. as a result of pipetting errors and stochastic effects). The effect of each of these "noise" factors was analyzed in relation to the "signal", which is the average signal ratio difference between tissues. Two loci with an average 10-fold difference in signal ratio values between semen and blood were selected. Different PCRs were performed on the same semen sample using 1 ng and 2 ng DNA as template. The ratios obtained from these amplifications were compared to each other and to ratios obtained from amplification of different blood samples. The differences between the ratios obtained in different PCRs and using different amounts of input DNA were more than one order of magnitude smaller than the differences between the ratios obtained from the different tissues (Figure 1D).

Example 3

Combined semen detection and DNA profiling

In this example categorization of the DNA is into one of two categories: semen and non-semen where non-semen refers to all tissues except semen. The categorization is performed together with DNA profiling in the same reaction.

The tissue identification procedure utilizes the same platforms used in standard STR profiling (i.e. thermocycler and capillary electrophoresis instruments) and is therefore readily amenable to integration with DNA profiling in a single reaction. This capability was demonstrated by generating an integrated STR profiling and semen detection assay. A STR profiling kit (Profiler Plus) PCR was supplemented with primers specific for two semen identification loci (Table 4 in Figure 10). These loci consist of a 70 bps amplicon which is efficiently amplified only in semen and a 95 bps amplicon which is efficiently amplified in all tissues except semen. The assay was tested on pure samples of semen, urine (male),

venous blood, menstrual blood, vaginal secretion, and saliva. The correct pattern of semen identification amplification was observed and the tissue identification algorithm correctly identified presence/absence of semen in all samples. Figures 3A and B depict complete electropherograms of semen and urine samples (obtained from the same individual), and Figure 3C depicts FAM channel data from all samples including the two semen identification loci and three profiling loci. The observed signal ratio was 25.04 in the semen sample, and 0.04 - 0.089 in the other samples. The combined assay was also tested on mixtures of semen and saliva at various ratios (Figure 3D). For each of these mixtures the percentage of the semen in the sample was derived from the signal ratio of the semen identification loci and compared with the corresponding percentage derived from the profiling loci (the semen and saliva samples were obtained from different individuals and could therefore be differentiated based on STR loci). The percentages derived were comparable, with a maximum difference of 10% (Table 5 in Figure 11). Moreover, the presence vs. absence of semen was correctly determined in all samples (including a sample that contained only 13% semen; Figure 3D). The STR profiles obtained from the samples using the integrated profiling-semen detection assay were identical to profiles obtained from the same samples using ProfilerPlus (without semen detection).

Example 4

Categorization of Aged and degraded DNA samples

The stand-alone tissue identification assay (described in example 1) was tested on two DNA samples extracted from 20-month old stains (blood and semen on cotton) as a proof-of-concept for application to forensic-type samples. Both samples amplified well, similarly to fresh samples (Figure 5), and the algorithm correctly identified the tissue source of both samples. In order to simulate degraded DNA samples, two additional DNA samples (blood and semen) were subjected to digestion with DNaseI under conditions that yield partial digestion. These samples were then analyzed with the combined semen detection and profiling assay. DNaseI digestion resulted in dropout of alleles in the larger STRs in both samples such that their profiles were partial. In the blood sample only D3S1358 and Amelogenin alleles were typed, and in the semen sample dropout of alleles was observed in FGA, D21S11, D18S51, D13S317 and D7S820 (Figure 6). The smaller tissue identification loci amplified successfully with the expected pattern, and the algorithm correctly identified the presence/absence of semen in these samples.

Example 5

Combined profiling and categorization of DNA

This particular assay employs both conventional profiling of a DNA sample (with the Promega PowerPlex16 kit), and categorization of DNA into categories of natural vs. artificial.

One aspect of this assay, therefore, is to check the natural nature of the DNA against several or all types of artificial DNA (see the Definitions section), and therefore contains an analysis of several different ratios. The combination of ratios makes it possible to therefore compare a single parameter various DNAs synthesized from various methods. A single parameter, for example representation ratio, can distinguish between a sample that contains the entire genome (natural DNA but also MDA synthesized DNA – as in this example) and a sample that contains only a subset (for example, only CODIS loci).

Two samples of DNA were analyzed – a natural DNA sample extracted from the blood of subject A, and an artificial DNA sample synthesized by in vitro multiple displacement amplification using a commercial kit (from a minute amount of subject A's real DNA as template).

The two samples were first profiled with a conventional kit (PowerPlex16) and analyzed using a conventional profiling software (GeneMapperID-X). The scheme of this procedure can be briefly described as performing multiplex PCR using the PowerPlex16 primer mix on a DNA sample; separating the amplification products on a capillary electrophoresis machine; and then analyzing the output data with, for instance, the GeneMapperID-X software.

For each sample, the end product of this assay is a profile. The profiles of the two samples are identical, as can be seen in Figure 9A,B. Furthermore, the GeneMapperID-X software determined the profiles of both samples to be single contributor (i.e. from a single person) profiles, with no anomalies. This shows that a profile obtained from artificial DNA can be identical to a profile obtained from natural DNA. See D. Frumkin, *et al.*, *Authentication of forensic DNA samples*, Forensic Sci. Int. Genet. (2009), doi:10.1016/j.fsigen.2009.06.009, which is incorporated herein by reference.

Both samples of DNA were subsequently also analyzed by a combined profiling and categorization assay based on the following scheme: digest a DNA sample with HhaI and perform multiplex PCR using the PowerPlex16 primer mix with the addition of primers for amplifying an additional locus, such as Hypo23; separating the amplification products on a capillary electrophoresis machine; and analyzing the output data with the capillary analyzer software, such as the software algorithm disclosed herein.

For each sample, the end product of this assay is a profile and assignment of category. As can be seen in Figure 9 C and D, the natural DNA was assigned as "natural", in contrast to the artificial sample which was assigned as "artificial".

In this example, the assignment of category was performed by analysis of three parameters:

SR1 (signal ratio of TPOX/D8S1179), SR2 (signal ratio of D3S1358/Hypo23), and RR (representation ratio of OCA2/D3S1358).

The threshold values for these parameters are provided in Table 6:

Table 6

	SR1	SR2	RR
Threshold for natural	≥ 0.9	≥ 10	≥ 2
Threshold for artificial	< 0.9	< 10	< 2

The observed values for these parameters, and the assigned category are provided in Table 7 (values that are above the threshold for natural DNA are shaded).

Table 7

	SR1	SR2	RR	Category
Natural DNA sample	1.23	17.54	3.31	Natural
Artificial DNA sample	0.06	2.06	2.47	Artificial

Assignment of the category was performed according to the following rule:

If all observed parameter values are above their respective thresholds for natural DNA, then assign a "natural" category, otherwise assign an "artificial" category.

Example 6

Combined profiling and categorization of DNA (MDA vs. non-MDA)

This particular assay employs both conventional profiling of a DNA sample (with the Promega PowerPlex16 kit), and categorization of DNA into categories of MDA (artificial DNA that was synthesized by Multiple Displacement Amplification vs. other (non-MDA) types of DNA.

Two samples of DNA were analyzed – a natural DNA sample extracted from the blood of subject A, and an artificial DNA sample synthesized by in vitro multiple displacement amplification using a commercial kit (from a minute amount of subject A's real DNA as template).

The two samples were first profiled with a conventional kit (PowerPlex16) and analyzed using a conventional profiling software (GeneMapperID-X). The scheme of this procedure is as follows:

DNA → multiplex PCR using the PowerPlex16 primer mix → separation of amplification products on a capillary electrophoresis machine → analysis of data with the GeneMapperID-X software

For each sample, the end product of this assay is a profile. The profiles of the two samples are identical, as can be seen in Figure 15A-B. Furthermore, the GeneMapperID-X software determined the profiles of both samples to be single contributor (i.e. from a single person) profiles, with no anomalies.

Both samples of DNA were subsequently also analyzed by the Capillary Analyzer software. For each sample, the end product of this assay is a profile and assignment of category. As can be seen in Figure 15C-D, the natural DNA was assigned as "non-MDA", in contrast to the MDA sample which was assigned as "MDA".

In this example, the assignment of category was performed by analysis of three parameters:

RR1 (representation ratio of vWA/D18S51), RR2 (representation ratio of CSF1PO/Penta_D), and RR3 (representation ratio of D21S11/D7S820).

A Combined Representation Score was calculated based on these three ratios as follows:

$$\text{CRS} = \sqrt[3]{R1 * R2 * R3}$$

The threshold value that was determined based on analysis of multiple MDA-samples and non-MDA-samples was 1.5, hence an unknown sample with a CRS value of above 1.5 is categorized as "MDA", while an unknown sample with a CRS value of below 1.5 is categorized as "non-MDA".

The CRS values of the MDA sample and non-MDA sample were calculated, and are as follows:

CRS value of MDA sample = 2.04

CRS value of non-MDA sample = 0.91

Based on the threshold value, both samples were correctly categorized.

Example 7

Stand alone semen detector kit

The assay employed by the kit accepts as input a DNA sample which then undergoes a biochemical procedure followed by signal analysis by a dedicated software. The output of the assay is the category of the DNA sample (semen or non-semen), and a statistical confidence level representing the likelihood that the outputted category is the true category of the DNA sample.

A semen detection kit of the present invention may comprise one or more of the following components:

1. A box that may comprise one or more of the following components:

Tube 1 comprising a 5X primer mix of:

0.6μM SD1f (AAGAGCCCATCAGGCAGGTC);

0.6μM SD1r (GTTTCTTGTCGAGCAGCACGTGGATGATG);

1.75 μ M SD2f (CTCCAGAACTGGAACTTCCTG);
1.75 μ M SD2r (GTTTCTTAACTTGGAGACGACGGCATC);
1.25 μ M SD3f (TGGAGGACAATGCCCTGGTG);
1.25 μ M SD3r (GTTTCTTGGCTTCACCTGCGACCGTCTC);
1.75 μ M SD4f (CCCTCCGAGTGGCCAGCAG);
1.75 μ M SD4r (GTTTCTGACCACTGCCGTGGGAATG);
1.75 μ M SD5f (CTTCTCAGCCAATGGGAAGAG);
1.75 μ M SD5r (ACGTAGAAGGACCCGAGGAC);
0.9 μ M SD6f (TACAGACAAATCACTCAGCAGC); and
0.9 μ M SD6r (GTTTCTTGTCTGACACTCGGTTGTAGGTATT)

The forward primers (*e.g.*, SD_f) are fluorescently labeled;

Tube 2 comprising a 10X reaction buffer (150mM TRIS-HCl, 15mM MgCl₂, 0.2mM each dntp, 2.5 μ g BSA);

Tube 3 comprising HhaI restriction endonuclease;

Tube 4 comprising a control semen DNA sampl;

Tube 5 comprising a control non-semen DNA sampl;

Tube 6 comprising a DNA ladder; and

a Material Safety Data Sheet (MSDS).

2. A paper detailing the following protocol, or similar instructions:

(a) For each reaction, combine in a 0.2 ml tube (not provided) the following ingredients: 5 μ l 5X primer mix, 2.5 μ l 10X reaction mix, 0.5 μ l HhaI endonuclease, 0.5 μ l DNA polymerase, 0.5 ng DNA, and DDW (distilled water) to a total volume of 25 μ l.

(b) In addition to the tested DNA, set up a positive control semen reaction using the supplied semen DNA, a positive control non-semen reaction using the supplied blood DNA, a negative control reaction using DDW instead of DNA, and a digestion control reaction using DDW instead of HhaI.

(c) Place the reaction tubes in the thermal cycler and run the following program: 37°C for 15 min, 95°C for 11 min, followed by 30 cycles of 94°C for 1 min, 59°C for 1 min, 72°C for 1 min, followed by a final extension step of 60°C for 45 min, and hold at 25°C.

(d) For each post-amplification reaction, mix 1.5µl product with 24.5µl formamide and 0.5µl fluorescent size standard in a new 0.2ml tube.

(e) Denature samples at 95°C for 3 minutes, then immediately chill on ice for 3 minutes.

(f) Run denatured samples in the capillary electrophoresis machine using the following parameters:

Module = GS STR POP4 (1 mL) F;

Inj. secs = 5;

Inj. kV = 15;

Run kV = 15;

Run C = 60;

Run time = 24 min

(g) Analyze output .fsa files in the TissueIdentifier analysis software.

3. “TissueIdentifier” analysis software implementing the tissue identification algorithm as described in the “Algorithm and Software” section above, with the following modifications:

In step 3 of the algorithm, the set of pairs of loci used are: {SD1,SD6}, {SD2,SD6}, {SD3,SD6}, {SD4,SD6}, and {SD5,SD6}.

A correction for incomplete digestion was performed using the signal ratio of {SD5,SD6}. SD5 acted as a digested control locus and SD6 acted as an undigested control locus.

Reference probability functions of semen and non-semen were obtained for {SD1,SD6},{SD2,SD6},{SD3,SD6},{SD4,SD6} using a set of 27 reference semen samples obtained from 27 different individuals and 86 reference non-semen samples (blood, saliva, urine, vaginal swab, menstrual blood) obtained from different individuals.

The assay is based on analysis of 6 genomic loci. Of these, 4 loci were found by the inventors to be differentially methylated in semen vs. non-semen tissues. Two of these four loci are methylated in all tissues except semen, and the two other loci are unmethylated in all tissues except semen. The assay also includes an undigested control locus that does not contain a HhaI recognition sequence (SD6) and is therefore expected to be amplified successfully regardless of the tissue type, and a digested control locus (SD5) which was found by the inventors to be unmethylated in all potential tissues.

The kit was tested on 27 semen and 86 non-semen (blood, saliva, urine, vaginal swab, menstrual blood) DNA samples from 95 different donors of different sexes, ages, and ethnicities. The algorithm correctly identified the source of all samples with a typical confidence level > 99.999999%. Figure 12 shows sample plots of 6 samples. The tested samples were the same samples used for obtaining reference probability functions. However, in order to avoid bias, in each specific analysis, the analyzed sample was not used for obtaining the reference probability functions.

THE CLAIMS DEFINING THE INVENTION ARE AS FOLLOWS:

1. A method for categorizing a DNA sample, the method comprising the steps of:

(A) exposing the DNA sample to digestion by at least one restriction endonuclease (where the restriction endonuclease is selected from the group consisting of at least one methylation-sensitive restriction endonuclease and/or at least one methylation-dependent restriction endonuclease) so as to generate restriction enzyme-treated DNA;

(B) co-amplifying at least two different genomic loci from the restriction enzyme-treated DNA in a single reaction mixture, so as to generate a specific amplification product for each locus within the at least two genomic loci, where at least one of the loci is a restriction locus;

(C) determining a signal intensity for each amplification product in the single reaction mixture;

(D) calculating signal ratios between the signal intensities of the amplification products for at least one predefined pair of loci within the at least two different genomic loci;

(E) providing one or more reference values representing ratios between the signal intensities of the amplification products of the at least one predefined pair of loci within the at least two different genomic loci in known DNA categories; and

(F) comparing the signal ratios to the reference values,

where the DNA category whose reference values best approximate the signal ratios is determined to be the category of the DNA sample.

2. A method as claimed in claim 1, in which the at least two different genomic loci comprise (1) two or more restriction loci and (2) an undigested control locus which lacks a recognition sequence of the restriction endonuclease, and in which step (D) comprises the step of calculating signal ratios between the signal intensities of the amplification products of the undigested control locus and each of the two or more restriction loci.

3. A method as claimed in either of claim 1 or claim 2, in which, in step (D), the at least one predefined pair of loci comprises one or more successive pairs of loci within the at least two different genomic loci.

4. A method as claimed in any one of claims 1 to 3, in which the DNA category is selected from the group consisting of tissue type, cell type, physiological condition, pathological condition, age, ethnicity, sex, methylation level, species, cell lines, natural DNA and artificially-

synthesized DNA, risk of developing a pathological condition, prenatal conditions, prognosis, propensity to respond to treatment, effects of cell line sub-culturing and response to medication.

5 5. A method as claimed in claim 4, in which the DNA category is a tissue type or cell type selected from the group consisting of blood, saliva, semen, epidermis, urine, plasma, hair, menstrual blood, vaginal cells and/or secretions, brain, esophagus, lung, stomach, heart, duodenum, liver, gall bladder, intestine, kidney, adrenal gland, urinary bladder, urethra, colon, testicle, ovary, uterus, muscle, tendon, ligament, fat, cartilage, bone, endothelial cells, uterine cervix, lymph, thyroid, pituitary gland, cerebellum, and breast.

10 6. A method as claimed in claim 5, in which the tissue type or cell type is blood, saliva, semen, or epidermis.

15 7. A method as claimed in claim 4, in which the pathological condition is selected from the group consisting of cancer, inflammation, an auto-immune disorder, a metabolic disorder, an infection, a degenerative disease, hormonal imbalances, a disorder characterized by abnormal DNA methylation, a neurodevelopmental disorder of ICF (immunodeficiency, centromeric instability, and facial anomalies), Rett syndrome, and fragile X syndrome.

20 8. A method as claimed in claim 4, in which the prenatal condition is selected from the group consisting of Prader-Willi syndrome, Angelman syndrome, Beckwith-Wiedemann syndrome, fragile X syndrome, Russell-Silver syndrome, Transient neonatal diabetes mellitus, Albright hereditary osteodystrophy, McCune-Albright syndrome, Familial nonchromaffin paraganglioma, Maternal and paternal UPD14 syndromes.

9. A method as claimed in any one of claims 1 to 8, in which the one or more reference values represent mixtures of DNA categories.

10. A method as claimed in claim 9, in which the mixtures of DNA categories are mixtures of semen and non-semen at various ratios.

25 11. A method as claimed in any one of claims 1 to 10, in which the methylation-sensitive restriction endonuclease is selected from the group consisting of AatII, Acc65I, AccI, AciI, ACII, AfeI, AgeI, ApaI, ApaLI, AscI, AsiSI, AvaI, AvaII, BaeI, BanI, BbeI, BceAI, BcgI, BfuCI, BglI, BmgBI, BsaAI, BsaBI, BsaHI, BsaI, BseYI, BsiEI, BsiWI, BslI, BsmAI, BsmBI, BsmFI, BspDI, BsrBI, BsrFI, BssHII, BssKI, BstAPI, BstBI, BstUI, BstZ17I, Cac8I, ClaI, DpnI, DrdI, EaeI, EagI, 30 EagI-HF, EciI, EcoRI, EcoRI-HF, FauI, Fnu4HI, FseI, FspI, HaeII, HgaI, HhaI, HincII, HincII, HinfI, HinPII, HpaI, HpaII, Hpy166ii, Hpy188iii, Hpy99I, HpyCH4IV, KasI, MluI, MmeI, MspAII, MwoI, NaeI, NarI, NgoNIV, Nhe-HFI, NheI, NlaIV, NotI, NotI-HF, NruI, Nt.BbvCI, Nt.BsmAI, Nt.CviPII,

Paer7I, PleI, PmeI, PmlI, PshAI, PspOMI, PvuI, RsaI, RsrII, SacII, SalI, SalI-HF, Sau3AI, Sau96I, ScrFI, SfiI, SfoI, SgrAI, SmaI, SnaBI, TfiI, TscI, TseI, TspMI, and ZraI.

12. A method as claimed in claim 11, in which the methylation-sensitive restriction endonuclease is HhaI.

13. A method as claimed in any one of claims 1 to 10, in which the methylation-dependent restriction endonuclease is selected from the group consisting of McrBC, McrA, and MrrA.

14. A method as claimed in claim 4, in which the DNA category is natural DNA or artificially-synthesized DNA, and where at least one of the genomic loci comprises a core short tandem repeat (STR) locus used for DNA profiling.

15. A method as claimed in claim 14, in which the STR locus comprises a human locus selected from the group consisting of D16S539 (SEQ ID NO. 1), D7S820 (SEQ ID NO. 2), D13S317 (SEQ ID NO. 3), D5S818 (SEQ ID NO. 4), CSF1PO (SEQ ID NO. 5), TPOX (SEQ ID NO. 6), TH01 (SEQ ID NO. 7), vWA (SEQ ID NO. 8), FGA (SEQ ID NO. 9), D21S11 (SEQ ID NO. 10), D8S1179 (SEQ ID NO. 11), D18S51 (SEQ ID NO. 12), and D3S1358 (SEQ ID NO. 13), Penta D (SEQ ID NO. 14), Penta E (SEQ ID NO. 15), and Amelogenin (SEQ ID NOs. 16 and 17), D2S1338 (SEQ ID No. 18), D19S433 (SEQ ID No. 19), ACTBP2SE33 (SEQ ID No. 20), D10S1248 (SEQ ID No. 21), D1S1656 (SEQ ID No. 22), D22S1045 (SEQ ID No. 23), D2S441 (SEQ ID No. 24), and D12S391 (SEQ ID No. 25).

16. A method as claimed in claim 14, in which one of the genomic loci is a locus known to be unmethylated in all tissues of natural DNA.

17. A method as claimed in claim 16, in which the locus known to be unmethylated in all tissues of natural DNA is selected from the group consisting of Hypo23, Hypo28, and Hypo33.

18. A method as claimed in any one of claims 14 to 16, in which step (D) of the method comprises calculating at least one of the following ratios between signal intensities of amplification products of the following loci:

further comprising the step of calculating at least one of the following ratios:

- (1) D3S1358/D18S51,
- (2) D3S1358/D7S820,
- (3) D3S1358/Penta_D,

5

.0

.5

20

- (4) D3S1358/TPOX,
- (5) D3S1358/FGA,
- (6) TH01/Penta_D,
- (7) D21S11/D18S51,
- (8) D21S11/D7S820,
- (9) D21S11/Penta_D,
- (10) D21S11/AMEL,
- (11) D21S11/TPOX,
- (12) D21S11/FGA,
- (13) D5S818/D18S51,
- (14) vWA/D18S51,
- (15) D5S818/Penta_E,
- (16) vWA/Penta_E,
- (17) D5S818/D7S820,
- (18) D5S818/Penta_D,
- (19) D5S818/TPOX,
- (20) D5S818/FGA,
- (21) D13S317/D7S820,
- (22) D13S317/Penta_D,
- (23) D13S317/TPOX,
- (24) D13S317/FGA,
- (25) D16S539/D7S820,
- (26) CSF1PO/D7S820,

- (27) vWA/D7S820,
- (28) D8S1179/D7S820,
- (29) D16S539/TPOX,
- (30) D16S539/FGA,
- (31) CSF1PO/Penta_D,
- (32) CSF1PO/TPOX,
- (33) vWA/Penta_D,
- (34) AMEL/TPOX,
- (35) AMEL/FGA,
- (36) vWA/D8S1179,
- (37) vWA/TPOX,
- (38) vWA/FGA,
- (39) D8S1179/TPOX, and
- (40) D8S1179/FGA.

19. A method as claimed in any one of claims 14 to 18, in which at least one of the genomic loci is TPOX (SEQ ID NO. 6) or FGA (SEQ ID NO. 9).

20. A method as claimed in any one of claims 1 to 19, in which the signal intensity is the amplification product's fluorescence level measured during capillary electrophoresis.

21. A method as claimed in any one of claims 1 to 20, in which steps (B) and (C) are performed by real-time PCR.

22. A method as claimed in any one of claims 1 to 21, in which step (B) is performed using fluorescently-labeled primers, and in which the signal intensity of each amplification product is its fluorescence level.

23. A method as claimed in claim 22, in which step (B) further comprises the step of hybridizing a fluorescently-labeled probe to each amplification product, and where the signal intensity

of said each amplification product is the fluorescence level measured after hybridization of the probe to the amplification product.

24. A method as claimed in any one of claims 1 to 23, further comprising the step of assigning confidence levels to potential categories of the DNA sample, where the confidence level reflects the likelihood that a particular DNA category indicates the categorical source of the DNA sample, and where the likelihood is calculated by:

(A) assigning a probability score to each comparison of a signal ratio to the reference values that correspond to the various categories of DNA, where the probability score is obtained by assigning a gamma probability distribution function to the reference values of each potential category; and where the probability score is equal to the value of the gamma probability distribution function assigned in step (A) at the observed signal ratio;

(B) for each category, calculating a category likelihood score, which is the product of the probability scores obtained in step (A); and

(C) normalizing likelihood scores by dividing the category likelihood score of each potential category by the sum of all category likelihood scores;

where the confidence level of each category is the normalized likelihood score for that category.

25. A method as claimed in any one of claims 1 to 24, further comprising the step of simultaneously performing DNA profiling with the DNA categorization.

26. A method as claimed in any one of claims 1 to 25, in which the steps of digesting and amplifying are performed together in a single vessel.

27. A method as claimed in claim 1, further comprising the steps of:

(A) providing a plurality of DNA samples;

(B) digesting each DNA sample with a methylation-sensitive and/or a methylation-dependent restriction endonuclease;

(C) performing PCR on each digested DNA to amplify at least two genomic loci, of which at least one is a restriction locus;

(D) for each DNA sample, determining the intensity of the signal of each amplification product;

(E) for each DNA sample, calculating signal ratios between the intensities of the signals produced by the loci; and

(F) calculating a distance measure of differential methylation between the DNA samples by performing a quantitative comparison of each of their respective signal ratios.

28. A method as claimed in claim 47, in which the distance measure is the sum of the absolute differences between the signal ratios of each of the DNA samples.

29. A method as claimed in claim 47, in which the distance measure is the square-root of the sum of squared differences between the signal ratios of each of the DNA samples.

30. A method as claimed in any one of claims 47 to 49, in which the distance measure between a subject sample and a reference healthy sample indicates the amount of medication required to treat a pathological state.

31. A method as claimed in any one of claims 47 to 50, in which the distance measure between a DNA sample obtained from cultured cells to a reference sample indicates the number of sub-culturing procedures that the cells underwent.

DATED: 6 November 2015
 NUCLEIX LTD, by its Patent Attorneys
 KNIGHTSBRIDGE PATENT ATTORNEYS

Figure 1

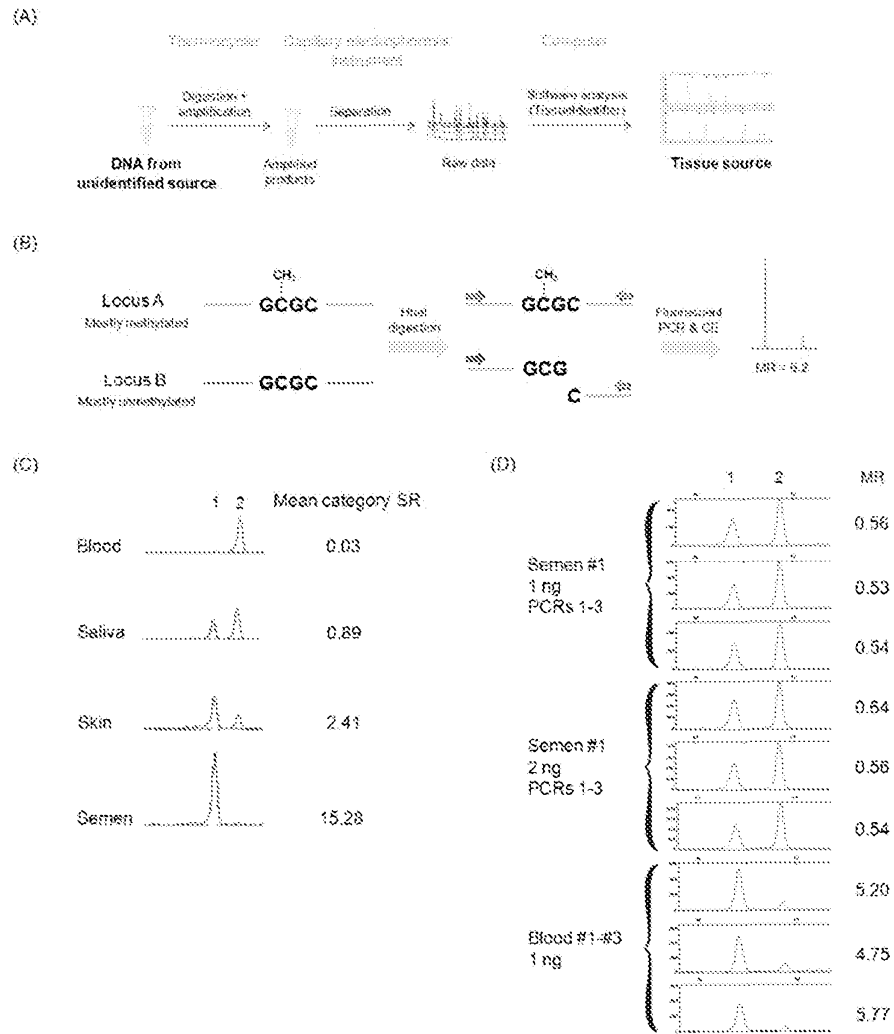


Figure 2

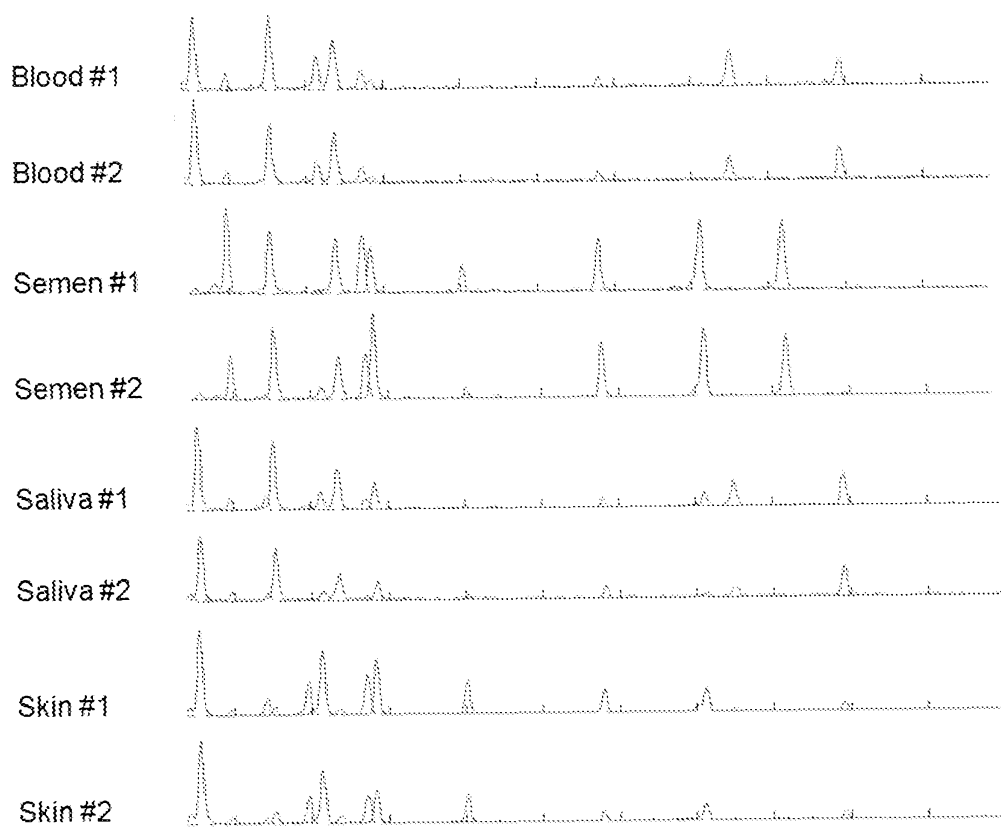


Figure 3

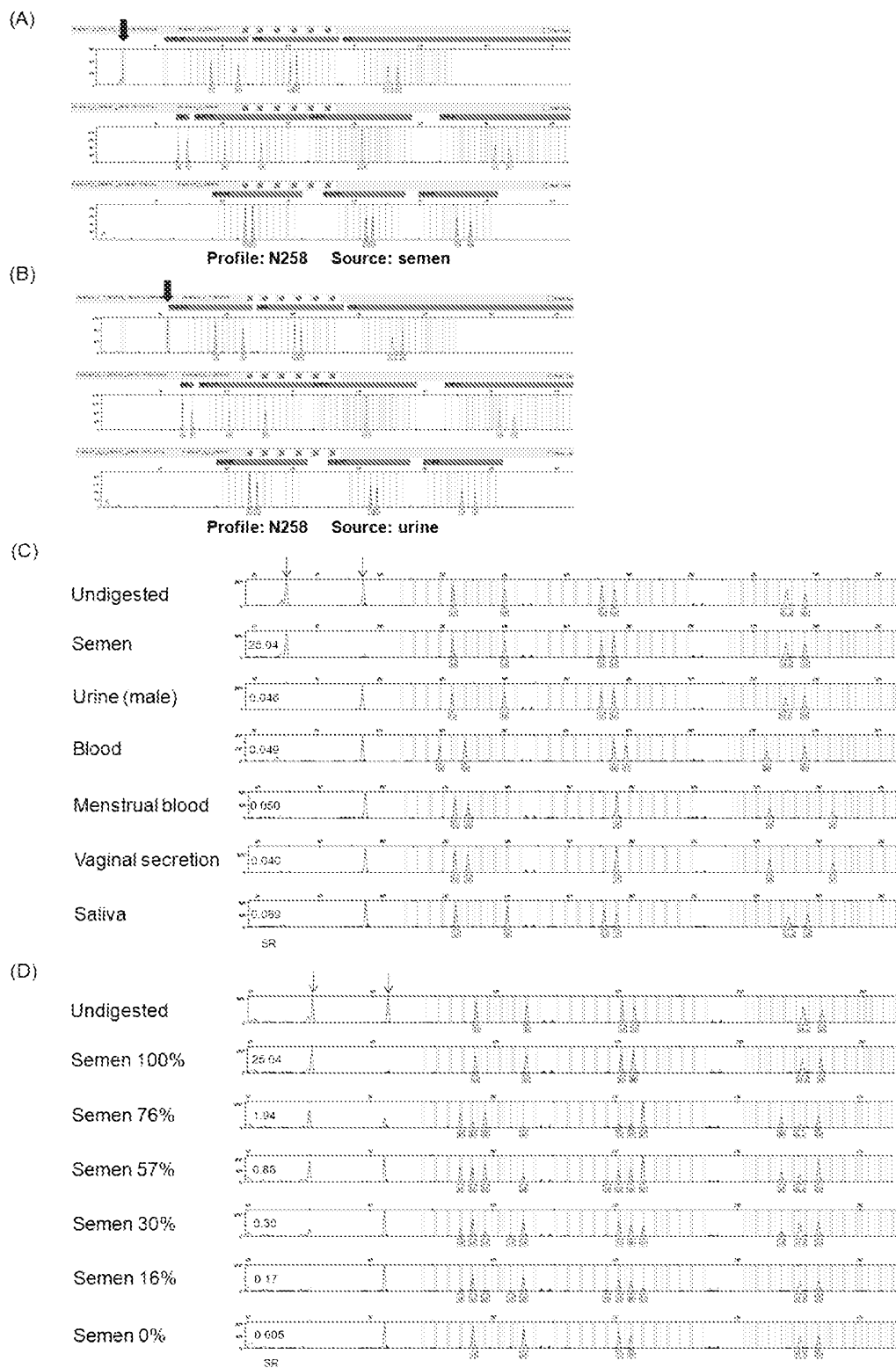


Figure 4

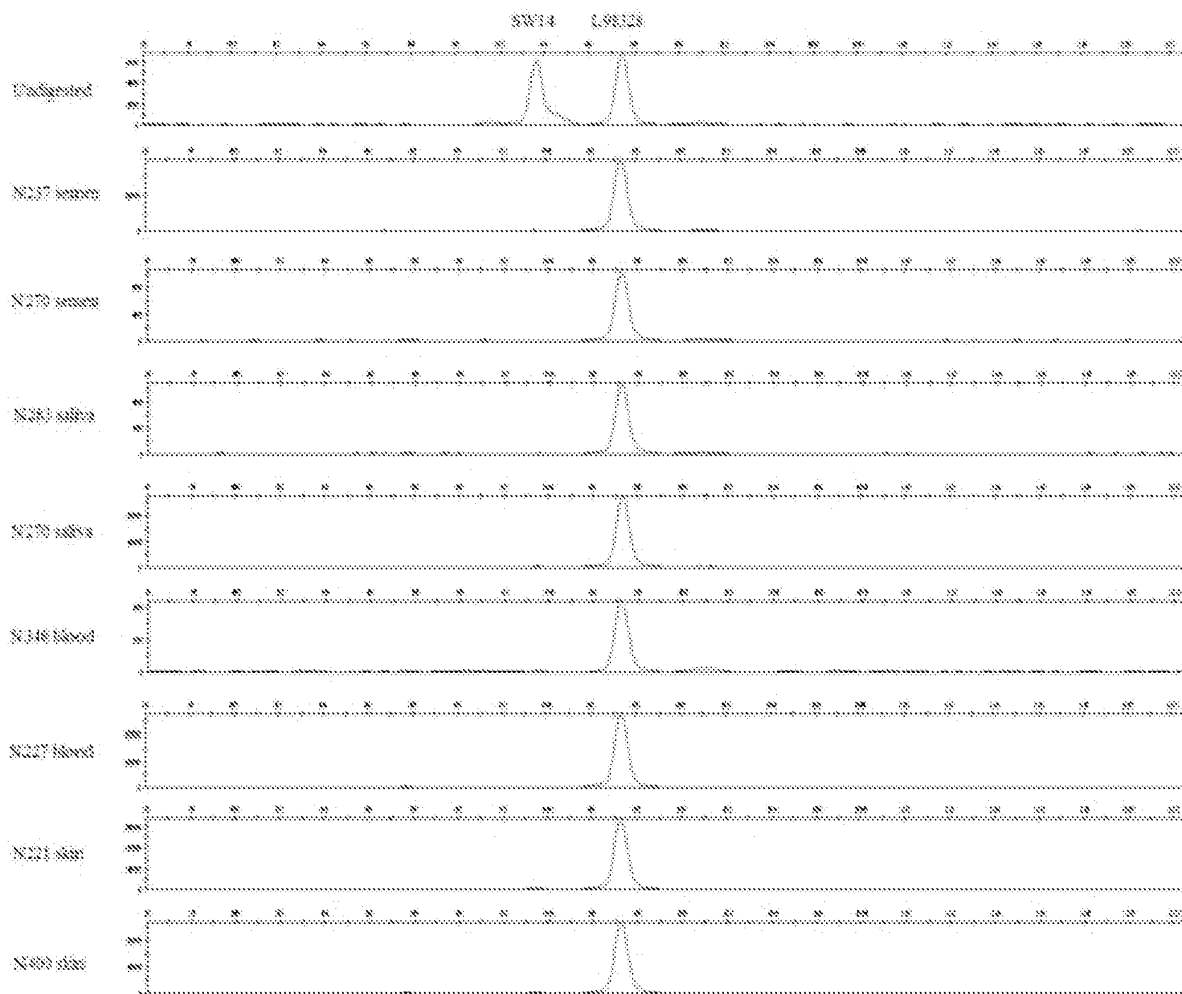


Figure 5

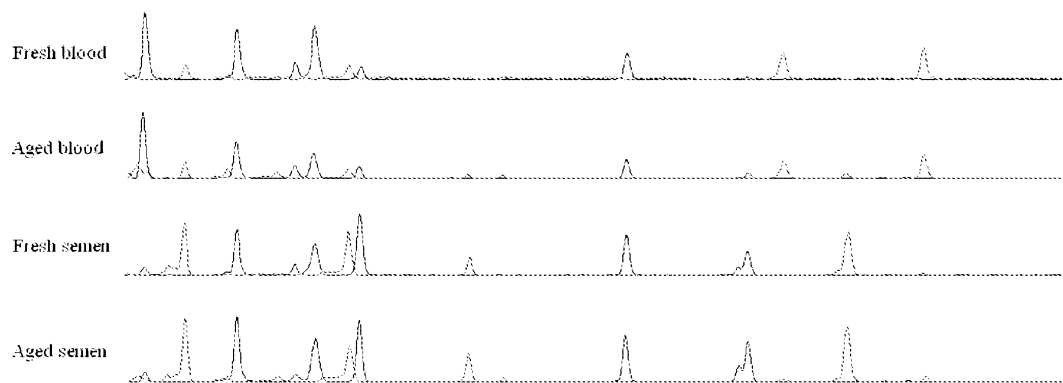
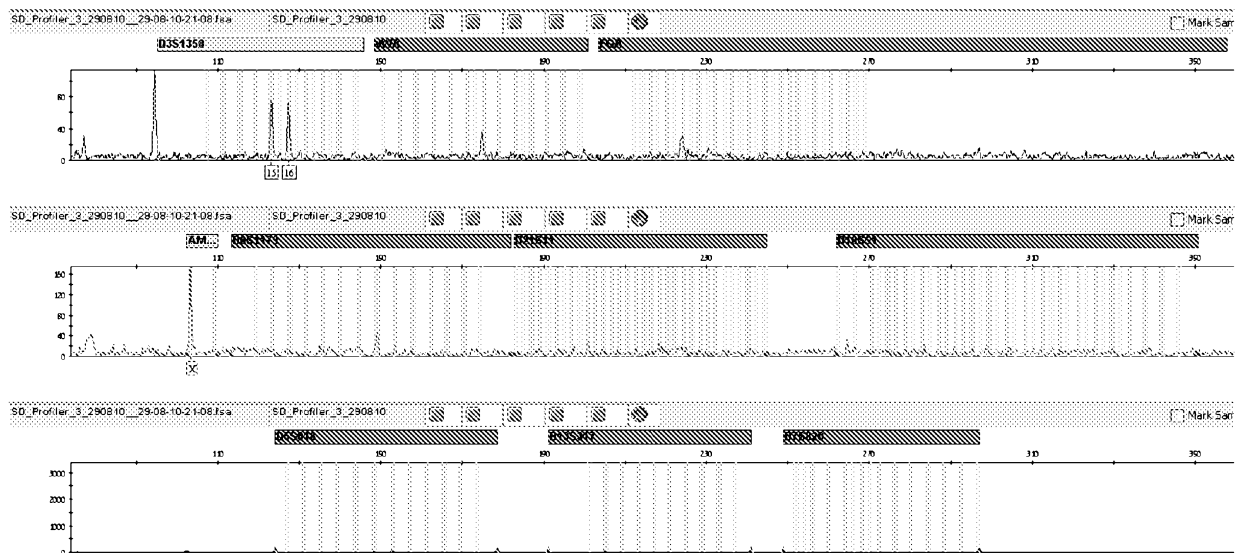


Figure 6

DNaseI-treated blood sample



DNaseI-treated semen sample

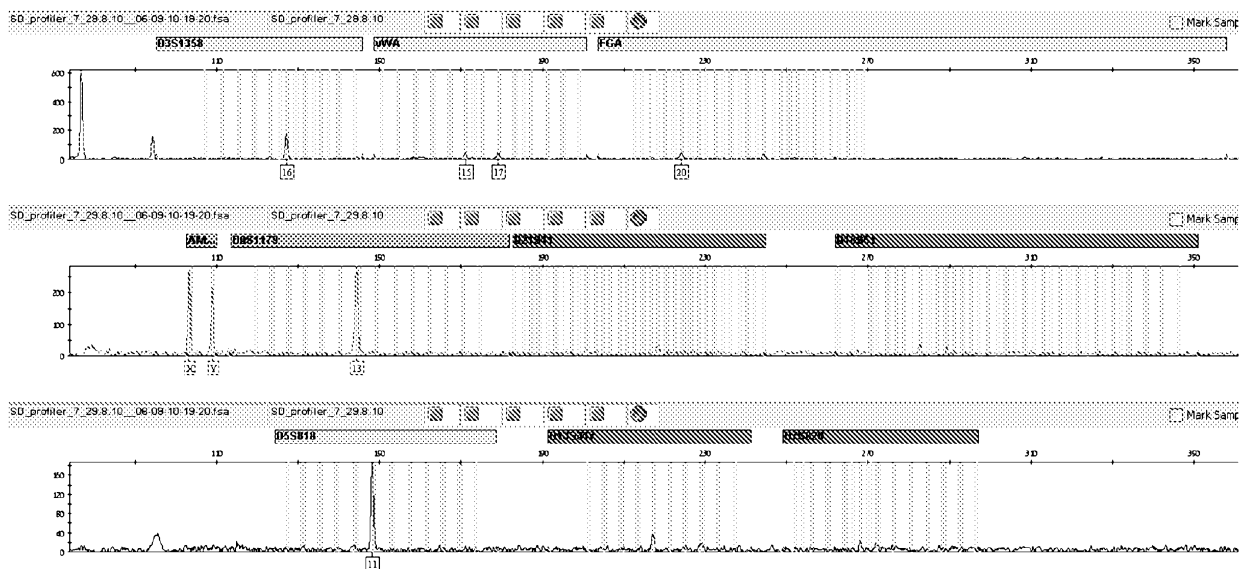


Figure 7

7/12

Table 1 – Panel of loci used for tissue identification assay

#	Locus	Location ^a	Forward primer	Reverse primer	Size ^b	Conc. ^c
1	L91762	Chr12 (73985697)	GCAGCAGGCCCGGGAGAAG (FAM)	AGCAGCTGTGCCGGGGCCAG	66	0.2µM
2	L68346	Chr3 (12175317)	CAGCAACAGACCCAGCTTG (JOE)	CACAGGCTCAGTCGCGGATC	70	0.2µM
3	L50468	Chr3 (55492965)	AGGAAACCTCAGTAGCAAAATTG (JOE)	GCGAGACTTAGGTGTGCATC	75	0.2µM
4	L14432	Chr22 (29866149)	CGTAGGCTGCGGTGAGCTC (FAM)	GATCCATGCCCGCTGGGATG	75	0.2µM
5	L4648	Chr1 (35815331)	CAGCCTAGACGTCAAGTTACAG (JOE)	ACGACCTCCGGATCCAACTG	80	0.2µM
6	L39664	Chr17 (53710330)	CCCAGCTGGTTGGACATGTTG (FAM)	CATTCTCTCGTGGACGCC	82	0.2µM
7	L30139	Chr17 (36996489)	GAGAAAGCGGAGGATGAGAC (FAM)	CCGATCTCTCTCCGTCCTG	84	0.2µM
8	L55429	Chr5 (1548043)	GCCTTCAGCAGGAAGTCCAC (JOE)	CCTGTGCTCACACAGACTC	88	0.2µM
9	L62086	Chr19 (40478538)	GTGCATGGTGTCTGTTACTTC (FAM)	GAAGCTCTCGTGGACTACTTG	89	0.1µM
10	L76138	Chr19 (3130217)	CAGCCTGCTCTTCACTGCAG (JOE)	AGAGGCCGATGAAGCCGTAG	100	0.2µM
11	L15952	Chr7 (2741305)	CTCCCTGATTTACGACAAGTTC (FAM)	GACAGTATGCTGATGCTTCTTG	117	0.2µM
12	L36599	Chr19 (4867642)	AAGGCAGAGTTCCTGCTGTC (FAM)	CGGATGCGAGGAGGATCCTAG	130	0.2µM
13	L26688	Chr17 (77844826)	CGGACCAGATTGCTGGTCAAC (JOE)	CGACCTTGCCAGATGTTTGAC	134	0.2µM
14	L81528	Chr19 (47395603)	AGCCTCATCCACACTGACCAG (JOE)	TCAGAGCTCTCCTATCTGGAC	141	0.2µM
15	L36556	Chr19 (50962118)	GCCAGGCCGTTGATGATGAC (JOE)	GAATATGGAGGCCCTGGGCAG	149	0.3µM

^a Genomic location of the examined CG according to UCSC version hg18 is provided in brackets following the chromosome number^b in bps^c The concentration refers to both the forward and reverse primers

Figure 8

Table 2. Signal ratios of L91762/L68346 and L76138/L26688				
	L91762/L68346		L76138/L26688	
	mean±std	range	mean±std	range
Semen	0.16±0.14	0.04-0.53	6.50±5.80	2.04-19.00
Blood	4.36±1.69	2.15-7.73	0.30±0.20	0.08-0.76
Saliva	7.10±3.30	3.36-15.38	0.56±0.41	0.16-1.54
Skin epidermis	11.40±4.20	4.60-18.28	8.41±5.58	2.1-19.56

Figure 9

Table 3. Loci used for assessing HhaI digestion			
Location*	Forward primer	Reverse primer	Size (bp)
Chr7 (141084738)	GCGAAGGAAGTGCTGGAGTC (FAM)	GTTTCTTGAAAAGGGCCAGACAC	94
Chr11 (86241803)	CAAAGTACTGGGGTTACAGGTG (FAM)	GGATGAACCTTTAAGACATCATC	97

* Genomic location of the examined CG according to UCSC version hg18 is provided in brackets following the chromosome number (in locus L98328, which does not have a HhaI recognition site, the location is of the beginning of the forward primer)

10/12

Figure 10

Table 4. Panel of loci used for semen detection					
Locus	Location ^a	Forward primer	Reverse primer	Size ^b	Conc. ^c
L68346	Chr3 (12175317)	CAGCAACAGCACCCAGCTTG (FAM)	CACAGGCTCAGTCGCGGATC	70	0.2µM
L16264	Chr2 (2734005)	GGACGAGTTAACTTCCTTAATTTC (FAM)	GTTTCTTCGCGGAACCTGGTTTAACTTC	95	0.2µM

^a Genomic location of the examined CG according to UCSC version hg18 is provided in brackets following the chromosome number

^b in bps

^c The concentration refers to both the forward and reverse primers

11/12

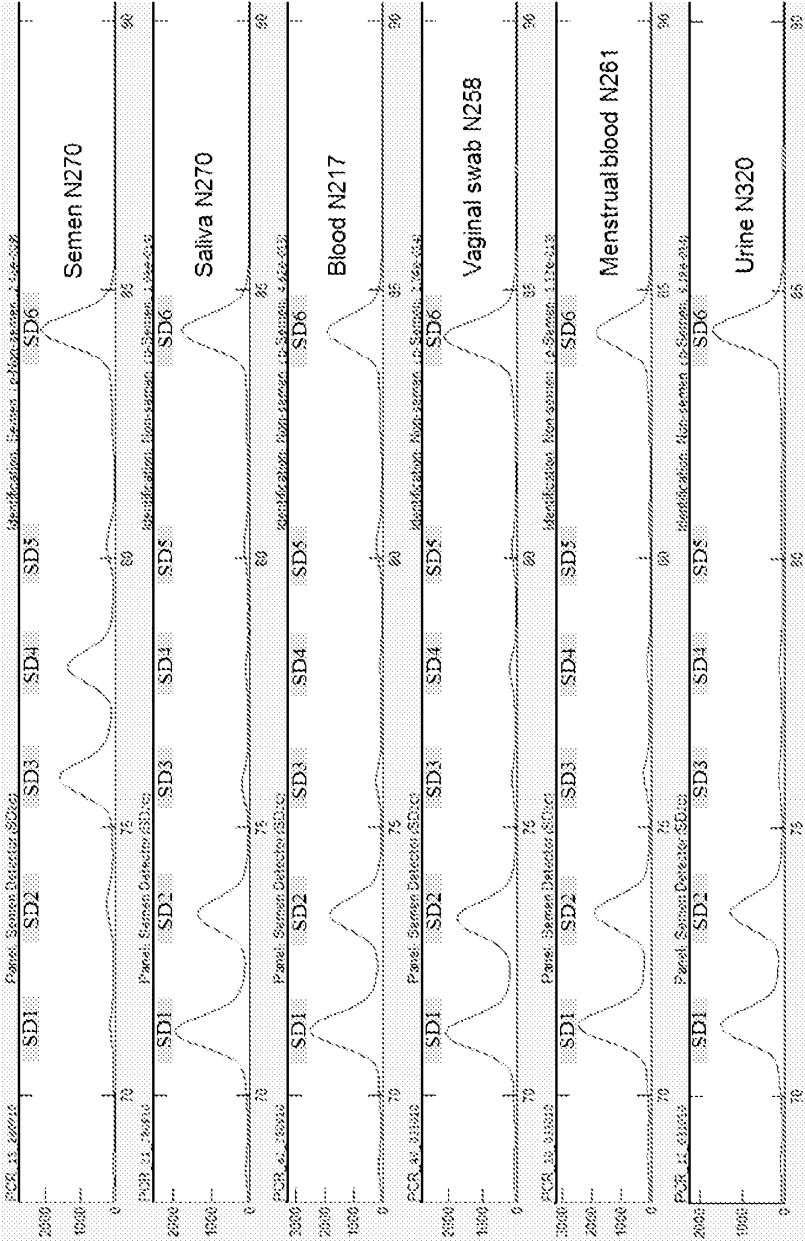
Figure 11

Table 5. Detection and quantification of semen component in semen-saliva mixtures			
Sample number	% semen in sample (according to STR typing)	% semen in sample (according to our assay)	Semen detected?
1	100	100	Yes
2	100	100	Yes
3	100	100	Yes
4	76	66	Yes
5	57	47	Yes
6	30	23	Yes
7	16	15	Yes
8	0	0	No
9	0	0	No
10	0	0	No

Each row represents a single reaction

12/12

Figure 12



SEQUENCES

Sequence 1: D16S539

```

1 CTCTTCTCAT TCCACAAGCT CTCCCCAAAA GACCCCATTC CTCCCCACCT TCAACCATCT
61 CTGGCAGGGA GGAGGGGGAA CTGAGAGGCT ACTTTTCTGAC CCAGGACCCCT AAGCCTGTGT
121 ACGGAGAGAG CATGAGCTGG GTGAGCTGCT TGCCAAGGAG TGGCATCTGC CCTCATCAGT
181 GGACACAAAA AGCCCCAGGG GTTAAGTGGC CATGGCTGCC CTCATGGCTG CACCGGGAGG
241 ATGACTGTGT TCCCACTCTC AGTCCTGCCG AGGTGCCCTGA CAGCCCTGCA CCCAGGAGCT
301 GGGGGGTCTA AGAGCTTGTA AAAAGTGTAC AAGTGCCAGA TGCTCGTTGT GCACAAATCT
361 AAATGCAGAA AAGCACTGAA AGAAGAATCC AGAAAACCA AGTTCCCAT TTTATATGGG
421 AGCAAAACAAA GGCAGATCCC AAGCTCTTCC TCTTCCCTAG ATCAATACAG ACAGACAGAC
481 AGGTGGATAG ATAGATAGAT AGATAGATAG ATAGATAGAT AGATAGATAT CATTGAAAGA
541 CAAAACAGAG ATGGATGATA GATACATGCT TACAGATGCA CACACAAACG CTAAATGGTA
601 TAAAAATGGA ATCACTCTGT AGGCTGTTTT ACCACCTACT TTAATAAAT AATGAGTTAT
661 TGAGTATAA TTAATTTTAT ATACTAATTT GAAACTGTGT CATTAGGTTT TTAAGTCTAT
721 GGCATCACTT TCGCTTGAT TTTTCTATTG ATTTCTTTTC TTTTCTTTTC TTTTGTGAGA
781 CAGAGTCTCA CTCTCACCCA GGCTGGAGTA CCGTGGCAGC ATCTTGGCTC ATTGCAACCA
841 CCACCTCCCG GGCTCAAGTG ATTATCTGCT CTCAGCCTCC CAAATAGCT

```

Sequence 2: D7S820

```

1 ATATGCTAAC TGGATGTGAA CAATTGTGTT CTAATGAGCT TAATATGAGT TTCATAATTT
61 GTGCATTTTG CTGTTAAAAA GCCAGAAAAC AAAACAAAAC AAAATACTGA AACCAGTGTG
121 AACCAAGATT ACACGATGGA AGGCATCAGT TTTCACACCA GAAGGAATAA AAACAGGCAG
181 AAATACCATA AGTTGATCCT CAAAATATGA TTGATTTTAA GCCTTATGAG ATAATTGTGA
241 GGTCTTAAAA TCTGAGGTAT CAAAAACTCA GAGGGAATAT ATATTCTTAA GAATTATAAC
301 GATTCACAT TTATCCTCAT TGACAGAATT GCACCAATA TTGGTAATTA AATGTTTACT
361 ATAGACTATT TAGTGAGATT AAAAAAAGCT ATCAATCTGT CTATCTATCT ATCTATCTAT
421 CTATCTATCT ATCTATCTAT CTATCTATCT ATCGTTAGTT CGTTCTAAAC TATGACAAGT
481 GTTCTATCAT ACCCTTTTATA TATATTAACC TTAATAAAT TCCATAGTCA GCCTGACCAA
541 CATGGTGAAA CCCCGTCTCT AAAAAAATA CAAAAATTAG CTGGATGCAG TAGCACATGC
601 CTGTAGTCCC AGCTACTCAG GAGGCTGGGG CAGGAGAACC ACTTGACCCA AGAAGCGGAG
661 GTTGCAAGTGA GCCGAGATCG CACCACTGCA CTCCAGCCTG GGTGACAGAG TGAGACTCCA
721 TCCTCAAGATA AAGAAATAAA TAAAAACAAA CAAACAAAAA AATTCCATAG GGGGTGAGGT
781 GCGGTGGCTC ATGCCTGTAA TCCCAGCACT TTGGGAGGCC GAAGCAGGTG GATCACTTGA
841 GGT

```

Sequence 3: D13S317

```

1 AATATGAATT CAATGTATAC AGAGAGAGAG AGAGAGAGAG AGAGAGAGAG AGACTTCTAC
61 AGAGCTCTAA GCATAATTGT GTAACCTCAA GCTCAGAGTG CCTAAGACCA GTACCAGGCT
121 GACTCATTTG AAAGCTGCCA TAGTAAGACT CTTCTGTTCA CTGCATTATT TATTGATGTA
181 TTGCAAGCAC TTAGTTACAT TTCTAGCATA TAACACATGA TCAATAAATA TTTTGACATG
241 AACAAATGGT AATTCTGCCT ACAGCCAATG TGAATATTGG GATGGGTTGC TGGACATGGT
301 ATCACAGAAG TCTGGGATGT GGAGGAGAGT TCATTTCTTT AGTGGGCATC CGTGACTCTC
361 TGGACTCTGA CCCATCTAAC GCCTATCTGT ATTTACAAAT ACATTATCTA TCTATCTATC
421 TATCTATCTA TCTATCTATC TATCTATCAA TCAATCATCT ATCTATCTTT CTGTCTGTCT
481 TTTTGGGCTG CCTATGGCTC AACCCAAGTT GAAGGAGGAG ATTTGACCAA CAATTCAAGC
541 TCCTCTGAATA TGTTTTGAAA ATAATGTATA TTAATGAATG TACAAATTTT CCCACTTGTA
601 CTTTCAGACT GTTATCTGTG AGTTAAAGCT CCTCCACTCT TTTTCTTACC CAAATAATAG
661 CATACTTTTT TCTGAGTATA TTTTGGGAAG AAGAGTTATT CAGTTATTGT TATATTTTAA
721 AAAATTCCTT ATACCAAAGT CTAATTGATC TAAGGCTATT CATTGAAACT TTCAGCATGC
781 TTAATAGCAG TC

```

Sequence 4: D5S818

```

1 CCCTCTGTGT AGCCTGGCTA TGTGCCACAT TGTGAGGTTT TCTCCCTCTA GCTACTTCTT
61 CCAGTTTCTT CTTTCAGGAT CCCAATTCCT TTCTCAAAGC ACTGGTGAAT AACTCCAAAT
121 ACTCCATCAT TTCAATTATC AGAGTAATAT AAGTCTTCTT TTTCTAAACC TCTCCCATCT
181 GGATAGTGGA CCTCATATTT CAGATGCTAA TAGGCTGTTG AGGTAGTTTC CTAAGCAAAA

```

```

241 AAGTAATTGT CTCTCTCAGA GGAATGCTTT AGTGCTTTTT AGCCAAGTGA TTCCAATCAT
301 AGCCACAGTT TACAACATTT GTATCTTTAT CTGTATCCCT ATTTATACCT CTATCTATCT
361 ATCTATCTAT CTATCTATCT ATCTATCTAT CTATCTTCAA AATATTACAT AAGGATACCA
421 AAGAGGAAAA TCACCCTTGT CACATACTTG CTATTTAAAT ATACTTTTAT TAGTACAGAT
481 TATCTGGGAC ACCACTTTAA TTAGAAGCTT TAAAAGCATA TGCAATGTCTC AGTATTTAAT
541 TTTAAAAATTA TTACATAAAT ATATACTCCT TTGAATTAGA AAATTACAAA TGTGGCTATG
601 TATTATTTTC TCCCATGTAT TTTCAAAATG AGGGGGTAAG CCAGACCCCT TCCCTCTCCC
661 ATGCCTAATA GCTCAAAGTT AAAGGTAAAG AAACAAGAAA ATATGGTGAA AGTTAACCAG
721 CTTCACTTCA GAGGA

```

Sequence 5: CSF1PO

```

1 ATTCAACACA TGAGGCACGG CCAGACCTAA ATGTCTCAGA GCCTGCTCCC ACTCCGATGA
61 GCTGCTGCCT TGCTTCAGGG TCTGAGTCCA GTGACTGCCA CTGCCTGCAC CCAATCACCA
121 TAGCCAGAGA CCTGGAGGTC ATCCTTATCT CTTTTCTCTT CCTCATCCCT GCATCTCAGA
181 CTCTTCCACA CACCACTGGC CATCTTCAGC CCATTCTCCA GCCTCCAGGT TCCCACCCAA
241 CCCACATGGT GCCAGACTGA GCCTTCTCAG ATACTATCTC CTGGTGCACA CTTGGACAGC
301 ATTTCCCTGTG TCAGACCCTG TTCTAAGTAC TTCCTATCTA TCTATCTATC TATCTATCTA
361 TCTATCTATC TATCTATCTA TCTATCTAAT CTATCTATCT TCTATCTATG AAGGCAGTTA
421 CTGTTAATAT CTTCATTTTA CAGGTAGGAA AACTGAGACA CAGGGTGGTT AGCAACCTGC
481 TAGTCCTTGG CAGACTCAGG TTGGAACACT GCCCTGGAGT GTGTGCTCCT GACCACCACG
541 AAGTGCCCTCC TCTGTACAAT CTGACCCCAT CACTCTCCTC TTTACAATGA CCTCCCAATA
601 GGTATAAGATG CAGTTATCTT TTCTCACTTT AAGACACCTT TACCTCCGGC TTCTGCCACC
661 TCCTCTGCTC CCCTGTGGCC ACTCCTCACA CCACTCCACA TCCCAGCTGT TGTCAAGTTC
721 TTTCAGTGTT CCAAAATGATC TATGTTCTCT TTGCCTTTGA GCCTTGCCATA TGTTCCTCCC
781 TCTGCCAGAA GCGCTGTTCC CCTTCCTTTC CCACCTTCTC GCCCCGCCAA CTCACCTTCA
841 TTCTTCCCAT CCCAGTTTAG AGGCCACCTT CTCGAAGCCT GGGTTGGGGG GACTCTTCAG
901 TGTTCCACAG ACACCTTGTG CTTCCCCCAT AATCACTGGG TGATCATTG

```

Sequence 6: TPOX

```

1 CCCAGCACAC ACCTTGCCCTC TGGCTGGGAC CCCCTTTGCT GCTGGCCCTG CTCAGGCCCC
61 ACAGCTTGAT CTCCCTCATGT TCCCACCTGCT GACTTCCCCA AGCTAACTGT GCCACAGAGT
121 GGGGGACCCC CTCCCGGCTC TCACAACCCC CACCTTCCTC TGCTTCACTT TTCACCAACT
181 GAAATATGGC CAAAGGCAAA AACCCATGTT CCCACTGGCC TGTGGGTCCC CCCATAGATC
241 GTAAGCCCAG GAGGAAGGGC TGTGTTTCAG GGCTGTGATC ACTAGCACCC AGAACCCTCG
301 ACTGGCACAG AACAGGCACT TAGGGAACCC TCACTGAATG AATGAATGAA TGAATGAATG
361 AATGAATGTT TGGGCAAAATA AACGCTGACA AGGACAGAAG GGCCTAGCGG GAAGGGAACA
421 GGAGTAAGAC CAGCGCACAG CCCGACTTGT GTTCAGAAGA CCTGGGATTG GACCTGAGGA
481 GTTCAATTTT GGATGAATCT CTTAATTAAC CTGTGGGGTT CCCAGTTCTT CCCCTGAGCG
541 CCCAGGACAG TAGAGTCAAC CTCACGTTTG AGCGTTGGGG ACGCAACAC GAGAGTGCTT
601 GGTGTGAGCA CACAGGAGGA GTCACGACAG AGCAGTGTA GAGCCGCCAC GTGGGTCCCA
661 CACAGGGGGA GTCACGACAG AGCAGTGTA GAGCCGCCAC GAGGGTCCCA CACAGGGGGA
721 GTCGCGACAC AGCAGTGTA GAGCCGCCAC GAGGGTCCCA CACAGGGGGA GTCACGACAC
781 AGCAGTGTA GAGCCGCCAC GAGGGTCCCA CACAGGGGGA GTCACGACAC AG

```

Sequence 7: TH01

```

1 TTACCCTTGG GGTGGGGGTG TAGGATGCAG CTGGGGCTGC AGTTCCAGGC CACGGAGAGC
61 CTGTGAGGCT GGGCCCCGGG GCGCCCTGGG GAGGGGATGC CTGATGGGGA GCCTGGTGGG
121 GGAGGGTAGG GGAGGGCGGG GGAGGACGGG GGAGGGCGCC CTGTGTCCCT GAGAAGGTAC
181 CTGGAATGA CACTGCTACA ACTCACACCA CATTTCAATC AAGGTCCATA AATAAAAACC
241 CATTTTAAAT GTGCCAGGGA GCCCAAGGTT CTGAGTGCCC AAGGAGGCAC CGAAGACCCC
301 TCCTGTGGGC TGAAAAGCTC CCGATTATCC AGCCTGGCCC ACACAGTCCC CTGTACACAG
361 GGCTTCCGAG TGCAGGTCAC AGGGAACACA GACTCCATGG TGAATGAATG AATGAATGAA
421 TGAATGAATG AGGGAATAA GGGAGGAACA GGCCAATGGG AATCACCCCA GAGCCAGAT
481 ACCCTTTGAA TTTTGCCCCC TATTTGCCCA GGACCCCCCA CCATGAGCTG CTGCTAGAGC
541 CTGGGAAGGG CCTTGGGGCT GCCTCCCCAA GCAGGCAGGC TGGTTGGGGT GCTGACTAGG
601 GCAGCTGGGG CAGAGGGAGG CAGGGGCAGG TGGGAGTAGG GTGGGGGCTG GGTGCAGCAG
661 CCGGGGACCT CTGGCCATCT TGGATTTTTT GGATGGATTT GTTTCACAT TCCGATCGTT
721 AAGATTCAAG ATGAAACAAG ACACAGAGAC CCACACGACC CCCGAG

```

Sequence 8: vWA

```

1 AGATGATAGA TACATATGTT AGACAGAGAT AGGATAGATG ATAGATACAT AGGTTAGATA
61 GAGATAGGAT AGATTATAAA TAGATACACA GGTTAGATAG ATTAGACAGA CAGATAGATA
121 CATACATAGA TATAGGATAG ATAACTAGAT ACAATAGAGA TAGATAGATA GATAGATAGA
181 TGATAGAGGA TAGATGATAA ATAGATATAT AGCTTAGATA GAGATAGGAT AGATGATAGA
241 TACATAGGAT AGATAGAGAC AGGATAGATG ATAAATAGAT ACATAGGTTA GATAGAGATA
301 GGACAGATGA TAAATACATA GGATGGATGG ATAGATGGAT AGATAGATAG ATAGATAGAT
361 AGATAGATAG ATAGATAGAT AGACAGACAG ACAGACAGAC AGATAGATCA ATCCAAGTCA
421 CATACTGATT ATTCTTATCA TCCACTAGGG CTTTCACATC TCAGCCAAGT CAACTTGGAT
481 CCTCTAGACC TGTTTCTTCT TCTGGAAGGT GGGAACTCTA CCTTATAGGA TCAGTCTGAG
541 GAGTTCACAA AATAATAAGG GCAAAGTGCC CGGCACATTG TAGGAGACTA GTAATGTCTA
601 TAAATGAGG GGCTTGAAGT AAATGATCCC TCTAGTTCTC TCTACTGCTA ACATTCTAAG
661 ACCTCCTTTA CATTAATTGT TCTCAAGCCA CATCTCCCTC CCCTACAGGA CTTCTATTTA
721 TTTCTGATCA ATTTCATGAG TACAAATAAG T

```

Sequence 9: FGA

```

1 ACTGAACATT TGCTTTTGAA ATTTACTATC TATGTACCGT TGGAAAATTT ACTTAATATC
61 TCTGAATTTT TTTTCTTCAA CTGTGGAGTG AGGAAAATAA TACCTACTTT TAGGTAGATG
121 ATGGATATAA CACTTTTCTC TGCATATAGT AGACACTCAG TGCATAACTA TCGCCTTCCT
181 TTTCCCTCTA CTCAGAAACA AGGACATCTG GGACCACAGC CACATACTTA CCTCCAGTCG
241 TTTTCATATCA ACCAACTGAG CTCTAACATT TTTCTGCAGA AGCTGGATAT GCTGTACTTT
301 TTCTATGACT TTGCGCTTCA GGACTTCAAT TCTGCTTCTC AGATCCTCTG AACTCGGTT
361 GTAGGTATTA TCACGGTCTG AAATCGAAAA TATGGTTATT GAAGTAGCTG CTGAGTGATT
421 TGCTGTAAAT TGCCAGCAAA AAAGAAAGGA AGAAAGGAAG GAAGGAGAAA GAAAGAAAGA
481 AAGAAAGAAA GAAAGAAAGA AAGAAAGAAA GAAAGAAAGA AAGAGAAAAA AGAAAGAAAG
541 AACTAGCTT GTAAATATGC CTAATTTTAT TTTGGTTACA GTTTAATCTG TGAGTTCAAA
601 ACCTATGGGG CATTTGACTT TTGGATAATG TTATGCCCTG CAGCCTTCCA TGAATGCCAG
661 TTAAGATGTC CTAATAGCAA TTAGTAATCC CAAAGAAATA TAGAAGAAGA ACTTTCTTTG
721 GAATTTTAAA GGTGTAATTT GGAGTTAAAA TAGTTGGTTT GATTGCAATT CAATTATTTT
781 ATAACATCCT TAATCAAGGG ACTTGAACAT ATTGGATTTT CTTACTGATG AGCTTTTCTT
841 TTTAATCTAT AGATTTGAAA TGGTTCCCTAA GCTGTTTGG GTCAACAGGA TCACTCACTT
901 GCCAGCTAGT GTTGCATCAC TGATTTTAAA TGTCAAGTGT TTGTG

```

Sequence 10: D21S11

```

1 GTTGGCTGGG GCTCAGAGAG AACAAAAAGG CAGAGGAAAA ACAATTTTCC CCTCTCACTT
61 CTGGAGATGG AACACTTTTC TTCTGCTTTT GGACATCAGA AATCCAAGTT CTCTGGCCTT
121 TGGACTTTGG GACTTGTGCC AGCACCTTCC TGGGTTCCTT GGCTTTTGGC CTCAACTGA
181 AGGTTACACT ATCAGCTTCC GTTGTTCTAA GGGCTTCAGA CTTGGACAGC CACACTGCCA
241 GCTTCCCTGA TTCTTCAGCT TGTAGATGGT CTGTTATGGG ACTTTTCTCA GTCTCCATAA
301 ATATGTGAGT CAATTCCCCA AGTGAATTGC CTTCTATCTA TCTATCTATC TGCTGTCTG
361 TCTGTCTGTC TGTCTATCTA TCTATATCTA TCTATCTATC ATCTATCTAT CCATATCTAT
421 CTATCTATCT ATCTATCTAT CTATCTATCT ATCTATCTAT CGTCTATCTA TCCAGTCTAT
481 CTACCTCCTA TTAGTCTGTC TCTGGAGAAC ATTGACTAAT ACAACATCTT TAATATATCA
541 CAGTTTAAAT TCAAGTTATA TCATACCACT TCATACATTA TATAAAACCT TACAGTGTTT
601 CTCCCTTCTC AGTGTTTATG GCTAGTAATT TTTTACTGGG TGCCAGACAC TAATTTTAT
661 TTTGCTAAGT GGTGAATATT TTTTATATCC TTAAAAATAT TTTTGAGTGT TGATCTGGGT
721 AAAGTTAAGT TCAATATTGG AAAAATATTG ATTCTTTTGA GGATAGTTAT CTTCTAATTA
781 GTCTACCTGT TGCCCCATAA ATGGCATGAT TTTCCACTCT GTG

```

Sequence 11: D8S1179

```

1 TACTACAGCA AGAGCGCTTG AACCAGATGT AGGGGAGATA GCAGCTGGAG AGCATAACAG
61 AGGCACTGAC ATGTGAGCAG CTAACGAGGC CTTTTACAAG ACATCTGTGA CCACACGGCC
121 AAGTAGAAGA AAGCCGTAA AAGCATCAAG GTAGTTAGGT AAAGCTGAGT CTGAAGTAAG
181 TAAACATTG TTACAGGATC CTTGGGGTGT CGCTTTTCTG GCCAGAAACC TCTGTAGCCA
241 GTGGCGCCTT TGCCCTGAGTT TTGCTCAGGC CCACTGGGCT CTTTCTGCCC ACACGGCCTG
301 GCAACTTATA TGTATTTTTG TATTTTCATGT GTACATTCTG ATCTATCTGT CTATCTATCT

```

```

361 ATCTATCTAT CTATCTATCT ATCTATCTAT CTATTCCCCA CAGTGAAAAT AATCTACAGG
421 ATAGGTAAT AAATTAAGGC ATATTCACGC AATGGGATAC GATACAGTGA TGAAATGAA
481 CTAATTATAG CTACGTGAAA CTATACCTAT GAACACAATT TGGTAAAAGA AACTGGAAAC
541 AAGAATACAT ACGGTTTTTG ACAGCTGTAC TATTTTACAT TCCCAACAAC AATGCACAGG
601 GTTTCAGTTT CTCCACATCC TTGTCAACAT TTGTTATTTT CTGGGTTTTT GATAATAGCT
661 GTGAAAGGAA AATAAAAACT TGGGCCGGGC GCGGTGGCTC ACGCTGTAA TCCCAGCACT
721 TTGGGAGGCC AAGGCGGGCA GATCTCAAGG TCGGGAGATT GAGACCATCC TGGCTAACAT
781 GGTGAAAACC CATCTCTACT AAAAATACAA AAACAAAAAA TTAG

```

Sequence 12: D18S51

```

1 CATGCCACTA AGCTGTACAC TGAAAAACGG TTAACATGAT AAATTTTATG TTACATACAT
61 TTTACCACAA TTTAAAAAAA TTATTAAAAA ATACTAACAA TAGGCCAAGC GTGATGGCTC
121 ACACCTGTAA TCCCAGCACT TTGGGAGGCT GAGACAGGTG GATCAATTGA GCTCAGGAGT
181 TTGAGACCAG CCTGGGTAAC ACAGTGAGAC CCCTGTCTCT ACAAAAAAAT ACAAAAATTA
241 GTTGGGCATG GTGGCAGGTG CCTGTAGTCT CAGCTACTTG CAGGGCTGAG CAGGAGGAG
301 TTCTTGAGCC CAGAAGGTTA AGGCTGCAGT GAGCCATGTT CATGCCACTG CACTTCACTC
361 TGAGTGACAA ATTGAGACCT TGTCTCAGAA AGAAAGAAAG AAAGAAAGAA AGAAAGAAAG
421 AAAGAAAGAA AGAAAGAAAG AAAGAAAGAA AGAAAGAAAG AGAGAGAGGA AAGAAAGAGA
481 AAAAGAAAAG AAATAGTAGC AACTGTTATT GTAAGACATC TCCACACACC AGAGAAGTTA
541 ATTTTAAATTT TAACATGTTA AGAACAGAGA GAAGCCAACA TGTCCACCTT AGGCTGACGG
601 TTTGTTTTATT TGTGTTGTTG CTGGTAGTCG GGTGTTGTTAT TTTTAAAGTA GCTTATCCAA
661 TACTTCATTA ACAATTTACG TAAGTTATTT CATCTTTCAA CATAAATACG CACAAGGATT
721 TCTTCTGGTC AAGACCAAC TAATATTAGT CCATAGTAGG AGCTAATACT ATCACATTTA
781 CTAAGTATTC TATTTGCAAT TTGACTGTAG CCCATAGCCT TTTGTGGCT AAAGTGAGCT
841 TAATGCTGAT CAGGTAAATT AAAAATTATA GTTAATTAAA AGGGCATAAA TGTTACCTGA
901 CTCAATAAGT CATTTCAATT AGGTCTG

```

Sequence 13: D3S1358

```

1 CTGGTTTTGG TGGAATTGAC TCCCTCTGTC ACAAACCTCAG CTTCAGCCCA TACCCTGAGC
61 CATGAGACCTA TCCCTCTAAT GCATTGTACT AGTCTCAGGG CTAATAACAA GGGAGAGGTG
121 TCAAAGGGCC AGTTCCACCT CCACCACCAG TGGAAGAGCT ATTCCCAGGT GAGGACTGCA
181 GCTGCCAGGG CACTGCTCCA GAATGGGCAT GCTGGCCATA TTCATTGGC CACTTCTGCC
241 CAGGGATCTA TTTTCTCTGTG GTGTGTATTC CCTGTGCCTT TGGGGGCATC TCTTATACTC
301 ATGAAATCAA CAGAGGCTTG CATGTAATCTA TCTGTCTATC TATCTATCTA TCTATCTATC
361 TATCTATCTA TCTATCTATC TATCTATCTA TGAGACAGGG TCTTGCTCTG TCACCCAGAT
421 TGGAAGTGCAG TGGGGGAATC ATAGCTCACT ACAGCCTCAA ACTCCTGGGC TCAAGCAGTC
481 CTCCTGCCTC AGCCTCCCAA GTACCTGGGA TTATAGGCAT GAGCCACCAT GTCCGGCTAA
541 TTTTTTTTTT TAAGAGATGG GGTCTCGCTG TGTTCCTCAG CCTTGCTCTA AACTCTGGC
601 CTCAAGTGAT CCTCCCATCT CAGCCTTCCA AAGTGCTGAG ATTACAGCAG AGGCTTTTAA
661 GTCAAAGCTT TCCCTGCTAG GACAAGCCCT AGTTAAAGTC CTGGAGCACT GGCCACTGCA
721 GCTGCACTTG G

```

Sequence 14: Penta D

```

1 CCTACTCGGG AGGCTGAGGC AGGAGAATCG CTTGAACCCA GGAGGGGGCG ACTGCAGTGA
61 GCCGAGATCG TGCCACTGCA CTCCAGCCTG GGTGACAGAG CGAGACTCCA TCTCAAAAAA
121 AAAAAAAAAA AAACAGAAATC ATAGGCCAGG CACAGTGGCT AATTGTACCT TGGGAGGCTG
181 AGACGGGAGG ATCGAGACCA TCCTGGGCAC CATAGTGAGA CCCCATCTCT AAAAAAAAAA
241 AAAAAATTT TTTTAAATA GCCAGGCATG GTGAGGCTGA AGTAGGATCA CTTGAGCCTG
301 GAAGGTCGAA GCTGAAGTGA GCCATGATCA CACCACTACA CTCCAGCCTA GGTGACAGAG
361 CAAGACACCA TCTCAAGAAA GAAAAAAAG AAAGAAAAGA AAAGAAAAGA AAAGAAAAGA
421 AAAGAAAAGA AAAGAAAAGA AAAGAAAAGA AAAGAAAAGA CGAAGGGGAA AAAAGAGAGAA
481 TCATAAACAT AAATGTAAAA TTTCTCAAAA AAATCGTTAT GACCATAGGT TAGGCAAAATA
541 TTTCTTAGAT ATCACAAAAT CATGACCTAT TAAAAAATAA TAATAAAGTA AGTTTCATCA
601 AAACTTAAAA GTTCTACTCT TCAAAAGATA CCTTATAAAG AAAGTAAAAA GACACGCCAC
661 AGGCTAAGAG AAAGTACTTC TAATCACATA TCTAAAAAAG GACTTGCTGC CAGATTAAAG
721 AATTCTTACA CATCAATAAG ACAACCAAT TAAAAATGGG CAAAAGATTT GAAGAGATAT
781 TTAACCAAG AAAACATATA AATGTGTCG GCGCGATGG TAATCCAGC ACTTTGAGAG
841 GCCGAGGCAG GCGGATCACT TGAGGTCAGG AGTTTAGGAC CAGTCTGGCC AACATGGTGA
901 AACCTGTCT CTAATAAAAA TACAAAATTT AGCTGGGTGT GGTGGCGTAA GCCTGTAATC
961 CCAGCTGCTC AGGAGGCTGA GGCAGAAGAA TTGCTTGAAC CTGGGAGGTG GAGGCTGCAG
1021 TAAGCG

```

Sequence 15: Penta E

```

1  CACATGTGGA  CATTTCCTAT  TTTCTCATAT  TGGTGGTATG  GCTCATTTAT  GAAGTTAATA
61 CTGGACATTG  TGGGGAGGCT  GTGTAAGAAG  TGTAAAGGG  GATCAGGGAT  ACATTCACTT
121 CTCTTTTCCT  TTGCTAGTTC  TGTGGTCTTA  AGCAAAGTAG  CCTCAAACAT  CAGTTTCCTC
181 TTTTATAAAA  TGAGGAAAAA  AATACTCATT  ACCTTGCATG  CATGATATAA  TGATTACATA
241 ACATACATGT  GTGTAAGTG  CTTAGTATCA  TGATTGATAC  ATGGAAAGAA  TTCTCTTATT
301 TGGGTTATTA  ATTGAGAAAA  CTCCTTACAA  TTTTCTTTTC  TTTTCTTTTC  TTTTCTTTGA
361 GACTGAGTCT  TGCTCAGTCG  CCCAGGCTGG  AGTGCAATGG  CGTGATCTCG  GCTCACTTCA
421 ATCTCCACCT  CCTGGGTTC  AGTGATTCTC  CTGTTTCAGC  CTCCAGAGTA  GCTGGGATTA
481 CAGGTGCCAT  CCACCACACC  CAGCTAATTT  TTTGTATTTT  AGTAGAGACG  GGGTTTCACC
541 ATGTTGCCCA  GGCTGGTCTT  GATCTCCTGA  GCTCAGGTAA  TACACCTGCA  TCGGCCTCCC
601 AAAGTGCTAG  GATTGCAGGC  GTGAATCACC  GCACCTGTCC  ACAATTTTCT  TGTTATTGGT
661 ACCCTTTCAT  GTTGGTAAAA  TGTATTTTAT  TTTCTCTTAT  CAAATAATTT  TCAATGCAAT
721 GAGAGTCAA  CTTTAAGCCC  AAGTAGACC  AGTAGTAAA  CTAAGGCTGA  AACCATTGAT
781 TGATTATTAC  CATATATTGT  CCTAAAATAT  TCGGCTTTTA  AAACATTTGG  TTTTCATTTT
841 CATGATAAAA  ATATGTAGCA  TTTTGTGACT  TTTAATTCAC  TTTGTAGAGT  TCTCAATCAT
901 TTCTAACACA  TGCTTGCCAA  TGACAAGCCA  TTTGTGAAAG  AGTTTGTGCT  GCTTAAAAAT
961 ATATGCAAA  GTAATAT

```

Sequence 16: AMEL X

```

1  AGGTCTCCTC  TTCTATACAG  CACATTTGTT  CAAACTAAAA  ACAGACCTCA  AGTATATTCT
61 GCACTATATA  GATTTTTTTA  AAGTAGCTTC  AGTCTCCTTT  AATGTGAACA  ATTGCATACT
121 GACTTAATCT  CTTCTCTCT  CTTCTCTTCC  TTCACCTCT  CCCTTCCTCT  CTCTTTCTAT
181 TCTCCTCCCC  TCCTCCCTGT  AAAAGCTACC  ACCTCATCCT  GGGCACCTG  GTTATATCAA
241 CTTTCAGCTAT  GAGGTAATTT  TTCTCTTTAC  TAATTTTGAC  CATTGTTTGC  GTTAACAATG
301 CCCTGGGCTC  TGTAAGAAT  AGTGTGTTGA  TTCTTTATCC  CAGATGTTTC  TCAAGTGGTC
361 CTGATTTTAC  AGTTCCCTAC  ACCAGCTTCC  CAGTTTAAGC  TCTGATGGTT  GGCTCAAGC
421 CTGTGTCGTC  CCAGCAGCCT  CCCGCTTGGC  CACTCTGACT  CAGTCTGTCC  TCCTAAATAT
481 GGCCGTAAGC  TTACCCATCA  TGAACCACTA  CTCAGGAGG  CTCATGATA  GGGCAAAAAG
541 TAAACTCTGA  CCAGCTTGGT  TCTAACCAG  CTAGTAAAAT  GTAAGGATTA  GGTAAGATGT
601 TATTTAAAC  TCTTTCCAGC  TCAAAAAC  CCTGATTCTA  AGATAGTCAC  ACTCTATGTG
661 TGTCTCTTGC  TTGCCTCTGC  TGAATATTA  GTGACTAAGT  GGTATA

```

Sequence 17: AMEL Y

```

1  TTATTCTCCA  ATATTTTGAA  ATGTGAATAT  TACAGTAATT  TCCCTTGTCC  AAATGAGAAA
61 ACCAGGGTTC  CAAAGAGAGG  AAATTATTTG  CCCAAAGTTA  GTAATTTTAC  CTAATCTTTA
121 CATTTTACCG  GATGGGATAG  AACCAAGCTG  GTCAGTCAGA  GTTGACTTTT  TGCCCTTTCA
181 TGAACCTTC  CTGAGCAGTG  GTTCATGAAT  GAATAAATTT  ACAGCCATAT  TTAGAGGAA
241 AGAGTCAATC  CGAATGGTCA  GGCAGGAGGG  TGCTGGAGCA  ACACAGGCTT  GAGGCCAACC
301 ATCAGAGCTT  AACTGGGAA  GCTGATGGTA  GGAACGTGTA  AATTGGGACC  ACTTGAGAAA
361 CCACTTTATT  TGGGATGAAG  AATCCACCCA  CTATTCCTTA  CAGAGCCAG  GGGACTGCTA
421 ATGCAACAG  TGATCAAAAT  TAGTAAAGAG  AAAAAATTACC  TCATAGCTGA  AGTTGATATA
481 ACCAGGGTGC  CCAGGATGAG  GTGGTAGCTT  TTATAGGGAG  GAGGGGAGGA  GAAGAGAAAG
541 AGAGAGGAAG  GGAGAGTGTG  AAGGAAGGGA  AGAGAGAGTA  AGAGATTAAG  TCAATATGCA
601 ATTGTTAACA  TTAAGAGAGA  CTAATAATTAC  TTTTAAAAAA  TCTATATAGT  ACAGAAATATA
661 TTTGAGGTCT  GTTTTTCTGT  AAAACAAGTG  TGCTATGTAG  GAGAGGAGAC  TT

```

Sequence 18: D2S1338

```

1  ACAAGGCACG  GAACTCACAC  CCAGCCTCTC  TCCATACAAC  AGAATATGGG  TTCTTGCGGA
61 GCTGGACTCT  GCAGGAGTCT  ATCTAATATG  GACTCTGTGT  CAATGACTCC  TGGGCCTCCT
121 CTGATCACCC  CATTAAGTTC  CTTGATTTGC  TTTGAGCCTC  AAATCTATGT  GACATCAATA
181 CGTTCATTTC  TTCTAGCAC  TTAGAAGTGT  TTCTTGTGTA  TACATTTGCT  GGCTTCTTCC
241 CTGTCTCACC  CCTTTTCTTA  CCAGAATGCC  AGTCCCAGAG  GCCCTTGTCA  GTGTTTCATG
301 CTACATCCCT  AGTACCTAGC  ATGGTACCTG  CAGGTGGCCC  ATAATCATGA  GTTATTCAGT
361 AAGTTAAAGG  ATTGCAGGAG  GGAAGGAAG  ACGGAAGGAA  GGAAGGAAG  AAGGAAGGAA

```

```

421 GGAAGGAAGG AAGGAAGGAA GGAAGGCAGG CAGGCAGGCA GGCAGGCAGG CAAGGCCAAG
481 CCATTTCGTG TTCCAAATCC ACTGGCTCCC TCCCACAGCT GGATTATGGG CCAGTAGGAA
541 TTGCCATTTT CAGGGTTTGG CTGTCACTGT AGTCAGGACC ATGAAGTCTT TAGGCACCTC
601 CACTCCACAC ACCCCCTGGT GAGAGCTCCC ATCTCCCTGT TCTGAAACAG CTCCCCAATA
661 TAGTACTGAT TCCGGTTAAA CTGGAACCCC TGCCCCTGCC CCTGCCCTG ATTTACATGA
721 GGACACTGAG GCCCAGAGGG GTAAAGTGAC TGCCAGGGGT CACACAGCTA GAAAGTGGCG
781 GTGCCAGAAC TGGAAGGAGG CCCTCATTC CAGTCACGG CTTTTCCATA GCACAGCCTT

```

Sequence 19: D19S433

```

1 ATGAACTGG ACACAGAAAC CAGACCCAG AGCACATACC GTATGAGTCC ATTGGTATGA
61 AGTTTAAAAA CAGATGGCAC TAGTCCAAAG GATTGGAAAGT TGGAATAGTG GTTACCAGGA
121 CTGGGGGGAG GAAGGGATGG TGGATGGTGA ACAAAGGAC CTTGGAGGGC TCCTGGGGTT
181 CTAGGAATCA ATCTTCCTTC TTTCCTCCTC TCCTTCCTTC CTCTTCTCT CTCTTCTTCT
241 GTTTTTATTT CAATAGTTT TTAAGGAACA GGTGGTGTG GTTACATGAA TAAGTCTTTT
301 AGCAGTGATT TCTGATATTT TGGTGCAACC ATTACCCGAA TAAAAATCTT CTCTCTTTCT
361 TCCTCTCTCC TTCTTCCTT CCTTCCTTCC TTCTTCCTT CCTTCCTTCC TTCTTCCTTA
421 CCTCTTTTCC TTCAACAGAA TCTTATCTG TTGCCCAGGC TGGAGTGCAG TGGTACAATT
481 ATAGCTTTTT GCAGCCTCAA CCTCCTGGGC TCAAGTGATC TTCTTGCCCC AGCCTCCTGA
541 CTAGCCAGGA CTACAGGAAT GTGCCAACAT GCCTGGCTAA TTTTAAAAAA TTTTATATAG
601 AGAAGAGGTC TCACTATGTT GCCCAGACTA GACTTGAAC CTTCCTCTCA AGTGATCTTT
661 CTGCATCAGT CTTCCAAAGT GCTGGGATTG CAGGCATGAG CCACCTCACC CAGCCTTAGA
721 AATGTTCTGT TTCTTGACCT GAGAGCTGGA TATACAGGAT TGCTCACTTT GTGAAAATTC

```

Sequence 20: ACTBP2SE33

```

1 GTACTTCAGA GTCAGGATGC CTCTCTTGCT CTGGGCCTCC TTGCCACAT AGGAGTCTTT
61 CTGACCCATG CCCACCATCA CTCCCTGGTG CCTAGGCTGC CCCACAATGG AGGGGAAGAC
121 GGCCTGGGGA GCCTTGCGCA TGCTGGAGCA GTTGTCGACG ACGACGAGCG CGGTGATAGC
181 ATCATCCATG GTGAGCTGGC GCGGGGTGCG GACGCAAGGC GCAGCGGCAA GGACAAGGTT
241 CTGTGCTCGC TGGGCTGACG CGGTCTCCGC GGTGTAAGGA GGTATATATA TATTTCTACA
301 ACATCTCCCC TACCGCTATA GTAACCTGCT CTTTCTTTCC TTCTTTCTT CTCTTTCTT
361 TTTCTTTCTT TCTTTCTTTC TTTCTTTCTT TCTTTCTTTC TTTCTTTTTC TTTCTTTCTT
421 TCTTTCTTTC TTTCTTTCTT TCTTTCTTCT TCTTTCTTTC TCTTTCTTTC TTTTCTTTTC
481 TTTTCTTTCC TTCTTCCTT TCTCTCTCTC TCTCTTTCTT TCTTTCTAAC TCTCTTTGTC
541 TCTTTCTTTC TTTCTTTTGA CGGAGTTTCA CTCTTGTCGC CCAGATTGGA GTGCAATGGC
601 ATGACCTCGG CTCACCTGAG CCTCCACCTC CCAGGTTCAA GCGATTATCC TGCCCTCAGC
661 TCCTTAGGAG CTGGAATTAC AGACGTGCAC CACCAAGCCT GGCTAATTTT TGTATATTA
721 GTAGAGACGG GGTTCACCT TGTGCGCCAG GCTGGTCTCG AACTCCTGAC CTCAGGTGAC
781 CCACCTGCCT TAGGCTCCCA AAGTCTGGG ATTATAGGCA TGAGCCACAG TGCCCAGCCT
841 TCTTTTCATT TAATACTATA GTAGTGTGAT CCTCTCTACC TATTACA

```

Sequence 21: D10S1248

```

1 TTCTGTTTTG CGTGTTTCC TAGTATGGTA CCTGGCCAAG GGCACACTAG ATCTTTGTCA
61 AGGTAATGAC TACTTTTTAT TAAATGCTTT CCATGTATCA AGTTCGTGTC CAAGCACTTG
121 ACATATATCA TTTTATTTTA TCCCGTGAAG TAGTTATTGG TATCTTCATT TACAAATAAA
181 AAAACAAGCT TAGTACTTAA CTCACGCTCT TGAACATAAT TATTGCTTTA AAGGTAGCTA
241 GGATTCTTAA TAGCTATTAT TACCAGGCA TGAACAATCA GTAAAAAGCA AACCTGAGCA
301 TTAGCCCCAG GACCAATCTG GTCACAAACA TATTAATGAA TTGAACAAAT GAGTGAGTGG
361 AAGGAAGGAA GGAAGGAAGG AAGGAAGGAA GGAAGGAAGG AAGGAAGGAA ATGAAGACAA
421 TACAACCAGA GTTGTTCTCT TAATAACAAG ACAAGGAAA AAGAGAACTG TCAGAATAAG
481 TGTTAATTAT AATATCCAGG GGTGGGATAC AGAGGTTTTA GCATCTGCTC TTTGCAAGC
541 ACTGCACCTA TTCTTGAGGA ATACCTGAGG GAAAAAGTAT GGTTCCTCAC AGGATCTAGT
601 TGGACTGGAA ATATGACATT CATATTGGAA TCCAGTGTCT TTTTCTGAAA AAGAGAGTTC
661 GTTCCAAGCT TAGCTCACAT GCAAGCTAAG ACAACCACTA GAAATTACTC TCCCCAGGGC

```

Sequence 22: D1S1656

```

1 GTCATGCCTA CAGTGTAACG GGAATTGACC AGGTAGGCGA CTTGAACTCC AACTGCAGGC
61 TATGGGGAGA CATGTGACAA TGCTAATCCC TTAGGCATTT ATTCACTGCA TTGCAGTTTA

```

```

121 AATGTCTGCC TTTCAGGCAT TTCAGAGATT ATGTCACCTA AAGAGGCAGG CTGGAATTCA
181 AAACGGCAAG CCAGGAAAGA GAGAAACCAT GTGATTCAC CGCAGCACAA AACTCGTTTA
241 GCAGCTGTAA GCGCCTGGTC TTTGTTTATT TTAAATTTC TTTCTTTCCC AATTCTCCTT
301 CAGTCTGTG TTAGTCAGGA TTCTTCAGAG AAATAGAATC ACTAGGGAAC CAAATATATA
361 TACATACAA TAAACACACA CACACCTATC TATCTATCTA TCTATCTATC TATCTATCTA
421 TCTATCTATC TATCTATCTA TCTATCTATC TACATCACAC AGTTGACCCT TGAGCAACAC
481 AGGCTTGAAC TTATATGGGG ATTTTCTTCC ATCTCTACCA CCCCTGAGAC AGCAAGACCA
541 ACTCCTCCTC CTCCCTTCTCA GCCTACTCAA CATGAAGATA ATAAGGATGA AGACCTTTAC
601 AATGACCCAG TTCCACTTAA TAAATAGTAA ATGTATTTC TCTTCCCTAT GATTTTCTTG
661 ATAACATTTT TTTTCTCTGG CTTATTTATT GTAAGAATAC AGTATATAAT ATAAATAATT
721 ATAAACATG TTAATTGGTT CTTTACGTTA TCGATAAGAC TTCTGGTCAA TGGTAGGCTA

```

TscI McrBC (half site)
752: Partially methylated BspDI ClaI

Sequence 23: D2S1045

```

1 GAGCCCAAGT TTAAACCCAG GCCCTCTGTG TCCCCCTACA GGGTGACTGC ATCTCCGAGT
61 CTGGCTTGT CATGCCTGAC AGAGGGCTGC CGAGTGAGCA GCTTAAGGCA TCCTGCCACT
121 GTGCAGCTGC CAACCCCTACA GCCCGGCAGC CCTGCGGGAG GAAGCTCTAG TGCAGGCCTC
181 TTAGGATCTG GGGTCCAGGA TGCTGATTTT AGGGCCGGGA CCTTGGGCAC CGTCCCTCTG
241 GTCTGCATAA GACCCACTAT GGGCAAACCT TAAACCTGAT CGTTGGAATT CCCCAACTG
301 GCCAGTTCTT CTCCACCTTA TAGACCTCTG CCTAGCCTTC TTATAGCTGC TATGGGGGCT
361 AGATTTTCCC CGATGATAGT AGTCTCATTA TTATTATTAT TATTATTATT ATTATTATTA
421 TTATTATTAC TATTATTGTT ATAAAAATAT TGCCAATCAT ACATTGCGGT GATCACTCAC
481 ACTGTGCCGG GCACTCTTGA GAGCACTTTA CATATATTGT CTCATTTAAT TCTCTCAACT
541 TGGGCACAGG CACTGTCACT ATTTCCATTC TACAGCTGAG GAGACTGAAG CACAGAGAGC
601 CTTAGGGACT TGCTTGAGGT CACACAGCTA AGAAATGGTG GAGCCAGGAT CAGAAACCAG
661 GCCACCTACA GAGCTCCCTG CAAGGGGAAC AGCATCCGGT TCCAGAGGCT GTGATTTTAT
721 CAGCTACACT GTGTGACTCC ATCTTCACAC TCTCCTGCCC CTCAAGAAGA CATATAACCT

```

Sequence 24: D2S441

```

1 ATGAAGAGAT GGTGAGGCGA GGTATGGGGG AAGGGGCGTG GAGCTTCCAT GTCTCCCTG
61 GGCGCCACCC TCCAGGAACC TCCACGTGTT CAGCTATACA GAAGCTTCCT GAACCCAGTC
121 CTCTTGGGGT TTGAGGGAAG CTTTATGACA TCAGCATTC TCCCTCCAGG GTATTAAATGG
181 AGCCCTCTCT GAAGAGATTC TTAAGACCCA CGGCCAGAAA GTTGGGTAAA GACTAGAGTC
241 CTGCCTTGGG GCAGGTGAAA GGAGTGCAAG AGAAGGTAAG AGAGATTCTG TTCTGAGCC
301 CTAATGCACC CAACATTTCTA ACAAAGGCT GTAACAAGGG CTACAGGAAT CATGAGCCAG
361 GAACTGTGGC TCATCTATGA AAATTTCTAT CTATCTATCT ATCTATCTAT CTATCTATCT
421 ATCTATCTAT CTATATCATA ACACCACAGC CACTTAGCTC CAATTTAAAA GATTAATCAT
481 AAACATTTGG GAAGGAGAGT GAAGATTTT GTGATGTTAA ATAAGAATGA TTATACTAAA
541 AACCAAAATA ATATGTTATT TATGGCTGGG TGTGGTGGCT TAAGCCTGTA ATCCAGAAC
601 TTTGGGAGGC CAAGGCTTGT GGATCACTTG AGCCAGAAAG TTCAAGACCA GCCTGGGCAA
661 CATAGGGAGA CCCTGTCTCT ACAAAAAATT TTAAATTTAG CTGGACATGA TGGCACGCAC
721 CCGTAGTCTC AGCTACTCAG GAGGCTCAG CCACTGCATT CCAGTCTGGG TAACGCACAC

```

Sequence 25: D12S391

```

1 GTCAGGAGTT CGAGACCAGC CTGGCCAACA TGGCGAAACC CTGTCTCTAC TAAAAATACA
61 AAAAAATTAG CTGGGCATGG TGGTGTGTTT CTGTAACCCC AGCTACTCAG GAGGCTGAGG
121 CAAGAGAATC GCTGGAACCC AGGAGGTGGA AGTTGCAGTG AGCTGAGATT GCACCACTGC
181 ACTCCAGTGT GGGCAACAGA GCGAGACTCT GTCTCAGAAA AAAAAAGAA TACATGAAAT
241 CAGAGAAACT CAAATTGTGA TAGTAGTTTC TTCTGGTGAA GGAAGAAAAG AGAATGATAT
301 CAGGGAAGAT GAAAAAGAG ACTGTATTAG TAAGGCTTCT CCAGAGAGAA AGAATCAACA
361 GGATCAATGG ATGCATAGGT AGATAGATAG ATAGATAGAT AGATAGATAG ATAGATAGAT
421 AGATAGACAG ACAGACAGAC AGACAGACAG ACAGATGAGA GGGGATTTAT TAGAGGAATT
481 AGCTCAAGTG ATATGGAGGC TGAAAAATCT CATGACAGTC CATCTGCAAG CTGGAGACCC
541 AGGACACTA GGAGCATGGC TCAGTCCAGG TCTAAAAGCC AAAAAACCAG GGAACTGAT
601 GGTGTAATTA TCCATCCCAG GTGGAAGGCC TGAGAACCTG GAGTGCCCTT GGTATAAGTC
661 CCAGAGTACA AAGACAGGAG AGCCTGGAGT TCTGACTTCC AAGGGCAGAA GAATGTGTCG
721 CAGCTCCAGG AGAGAGAGAG AAAGAAATTC TTCTCTCCGC CTTTTGATTC TATCTGGGGG

```

Sequence 26: ADD6 (from ncbi accession NT_022135)

```
16664701 cttgaacctg agaggcagag gttgcagtga gccgagacca tgccattgca ctccagcctg
16664761 ggcaatagag taaaactcca tcctcccgcct ccaaaaaagt agacaacgtc catgagggtga
16664821 tgaggaaggg gttatcgtgt gttgcttgct gagaacagga cccccagact caccgtgtcg
16664881 acgccggcca gcagcatctc agtcacgttg gcgtagatct cctgcagcgt cagagcctgg
16664941 ctaaggaaga ggtatgtgag aagtcccccg ctcaccctcc ggcctcggtc catttggtac
16665001 tgtatgtccc tcaacttggt gtcaacatga atttggcctg tttgaaaaca gtatttcttt
16665061 tgaaggaggt ttgggttgag aatcatcttt tcagtctcaa agccctctgt cctcccagta
16665121 gcttaactaa accagtggca ggtgacagag ggtaaggaaa cccaatttat ctaacgtcaa
16665181 cctgggagtt tcactcatat acttgcttat gtaaatgaat gaaaagttaa aagacaagct
```

Sequence 27: ADD10 (from chromosome 17:3477839-3478292)

```
1 AGAGAGCCCC GGAGACAGGC AGAAAGGAAG GCATGTGACC GGATCACAAT CATCAGCTCT
61 CTGCTGTCCT CTTTGGGAAG GGTTTTAGTA TTAAGAGGAC ATTTATTCTC ATTAATGCAA
121 AATTAAGGAG TTTTAAAGC TTTTACAACC TAGACTCCCT CTGAGAGGTT AGCCTTGACA
181 CCCTAATCGC CTCTGTCTCC CGCCACTGCT CGGTGCCAAG CAGCTCCAC GGCCCCGGCG
241 GGTCTGATGA TAGCCGGACA GGAGGGAGGA AGGGAGGAG GAAGAGCCTG CATCAGCTCC
301 TACGATTGCC CAGCCCCATC CTGGGAGTGA TTAACGGTG CATCACCATA TGCCAGTCCC
361 ACTGACAGGC AGGTCACCGT GCACTTCAGG GCACTCTAAA TTGCCGACTC TCCATGTAGA
421 GAGGGATGAA TCCAATATTG AAATCCTCAT AACTACAGCC CCCCAGAGTA GCCGTCCATC
481 TCTGCTTAA AATGTTGATC TGTAGTAAAA TGTGATTTT GTTGAAGCTG AGTGATG
```

Sequence 28: ADD17 (from chromosome 1: 50149332-50149574)

```
1 TTGAACCTAG GAGATGGAGG TTGCAGTGAG CTGCGATCAT GCCACTTCCC TCCAGGCTGG
61 GCAACAGAGC GAGACTCCAT CTCAAAAAA CAAAAAGAA AACCAACCTT TTGAATGTAG
121 GGGAACTTT TCAAAGGATA TCTAGTTTTC AATTACAGTA AACTTGTGGA AGGGAGGTTT
181 AGAGTTGAGA TTGAGATTAT AGATTTTGCT GATGATAAAC CATGAGTTCC AGAGGACATA
241 GTAGACTATT CTGGGCAGTT ATACAGGGGT GGATGGAATG TGGGAGTGGG GTTGTATAGT
301 GCCATAAAGA AATGAGAGTC CGGATTAAAA ATAATGAGCT GGAATCGCGA GCCTTTTGTA
361 ACTGAAATAA ATAGAAAAAT AAGAAATACA TTATTTCTGT GATTGTGAG AGGAAGAAAT
421 GGTGGAATC TTGTGAGAAG CACACTGAGC TCTAGCACA CCTCTTCACT CCTACAGATG
481 TTGGAATAAA CGGCAGGCAA GTTCAAAATC ACATATAGTC ATTATTGCAA GATAGTTCTA
541 TGGATATAGA TACTACATAC AATATAAATC ATGCTCATTG AATGGTTCAG TGGAACTAC
601 TCTGAACCT
```

Sequence 29: Hypo23 (from ncbi accession NM_004907)

```
123161 GAGTTTGGGA AGGGTATTTT AGGGGGGAAT AACTTTTGAG TTCCCAGCGT GCGGGGGAAG
123221 GGCGGGACGG GAGGGGTGTC CAAGGCCTGA GAAGATCAGT GTGGGGCAGG GGTGAGGAAT
123281 AACCTGGGAG GGGGCCCTGT ATGGGGGAAA TAATTGGGAA GAGGAGAGAT GGGATGAAGG
123341 GGGCCTCAGC GGGTCGTCTC CTGTGTATGC AGGGTCGTTC TGCAGCGTCT CTGGGAGATG
123401 GCGTCCCTGG GAGCCCTCAG GTCGCCCCCTA CCGCTGCGG GGTGCTTTCC TGCGCTCAGC
123461 CCTTCCTGGC CCCTGGAGGG AAGGAAGTGA AACTCTCCTC TTCCCCCACC CGGCTGGAAT
123521 GCGAGTCAGG AAGCCTGGGG CTCCAGCCTG CTCCGGCTGC CCGGTTCGGG GATGGGGAGG
123581 GGCGTGGCCG GAGCGCAAAG CCCC GCCCCT CCGCGCCCC CCCC CGGAAG CCCC GCCGCC
123641 GGCCGCTAAG GCGATCACGG GCCCTGTCTT AATATGGGCA ACCGGAAGCG GCCCGCGCGA
123701 CTGCCCTACG TCACTCCGTC CAAATTTAGT TGTGGAAGTC AGCGGGCGCT GGTGGCGGGA
123761 AGCGCGCGCG AGCCAGTGCG GGCGGAAAGG GGGCGGGGG CGCACCACC CTAAAGGGC
123821 CCGCACCAGG AATGAATGGA GCCATTCGAA CAATCTGCA TCCTATTTT GGAGGAAGTG
123881 GAATTAGTAT TT
```

Sequence 30: Hypo28 (from chromosome 5: 85949232-85949719)

```
949232 TTTATTTTAA AAAAGAAAAG AAAGAAGAGA AAAGGGATGG GTTTATTGTC CTTTCAACA
949292 GACTAGAGTA TACGGGGTGA AACTGCTTCA CTTGATTCAA TAAATCGTT TCCGGTAACA
949352 GGCCCCAGGA ATCCTAGACC TAAGCCTGGC GCGAACTAC ATTTCCACA ATCCTTCGGG
949412 GGCTGATAAG GCTCCGCAAT GGTCTGAAC ACAATPCCA CAATCCAGG CGATTTCCGC
949472 TTTGTCGCGT TTCCTCAAG CTCCGCCCA TTTCCCATCT TTCTTTTCAG TCCTTGCGCA
949532 CCGGGGAACA AGTCTGTGAA AAAAAAGGTC TTGGTGAGGT GCCGCCATT CATCTGCTC
949592 CATTCTCTGC GCCTTTCGCA GAGCTTCCAG CAGCGGTATG TTGGGCCAGA GCATCCGGAG
949652 GTTCACAACC TCTGTGTGCC GTAGAGCCA CTATGAGGAG GGCCCTGGGA AGGTTAGTGT
949712 GTAAGGGG
```

Sequence 31: Hypo33 (from chromosome 6: 34211067-34211314)

```
211067 ACCCTCATTT CACATTTTAC CCCTTCCTCA AAATGCTCCC TTCATATTAC CTCCTCAGAA
```

211127 ACCAAGAATA TGGCTACTAA TTCTCCCTGG CCCCATGCTG CAGGTGAACC GGTAGCCCAG
 211187 AGGTATCACA TAATTCTCCC AAAGTCACAC AGCAAATCAA GATGCATCCA GGAAGTAGAAG
 211247 CCATGTGAGC CACACTGGGA AGCCCCAGCG AAGCTGACAG AAAGTTTCAT AATACCACCC
 211307 TCTCCCCCT

Sequence 32: OCA2 sequence: (from chromosome 15: 25276967-25277446)

6967 aaacacccca gtctgaaaat aaccatagtt tgttgctctt acgagtgaat atgctatttc
 7027 atacacgaag ctttgccttt cagcacccaa gatttaagga taattatgga tgaatattat
 7087 ggattcattt taaatccttt ggcaaatctg ctctgggggc ttctctgtca gaaggtctct
 7147 ccttcccaac tctaagaac gttatttcta tgcaaatgct gctgagtcaa gacggggagg
 7207 gaagtgacga gagaagggt ggtggcatgg tcagtaagtc atgagggtga gattaggggt
 7267 gacacactgc ttgccaacgt aggagaaggc tctgccccta cctagcaggt ctgatggaag
 7327 ccccttattc cgtccttctt gcgggggtcc accgagatcc aaaaaggaat gctgtgtagg
 7387 agcacatgat atgtgataaa tgagagaaaag gtcaaacatt taaggaacgc ccagagaaaag

Sequence 33: SD1 (chromosome 11: 72085699 - 72086161)

72085699 CGCTGGGCGC GAGTGAAGAG CCCATCAGGC AGGTCGCGCA GGAAGCGCTT GAGCGCCGAG
 72085759 GAAACATCAT CCACGTGCTG CTCGCCCTCC TTGAGGTGCA CAGAGCGCGC ATCCTGCCGC
 72085819 AGGCTCTCCA GCAGCCGCTG TGTCTTCGAT GTCTGCCAC ACTTGCGGTA GATGCCCTCG
 72085879 GAGTCAAGGC CTAGGGAGGG GCGGGGCCAA GCGTTCGGGG CCTGAGGCAT AGAGTCATGG
 72085939 GCGGGGGCCG CGCAGCTCTG GGGCGGGAGG CGGCTCTCCG GGAGGGGCGG GGCTGGCACC
 72085999 CTAGGGGCGC GGCTCACC GCCTGCGTAT GTAGTCCACA CAGCGGTACA CGATCACC GG
 72086059 GATATCCGAG TCCCCAAGCT GCTGCTCCGA CAGCGTGTCC CCCATGCTGG CGGCTGCTTT
 72086119 CTGGATGGCC CCCAGCCAA CCATGAAGTC CAGCCGCGCG TCG

Sequence 34: SD2 (chromosome 4: 25286907 - 25287230)

25286907 CGATCCCGTT CACTCGCCTG CCCATCCGCA TGGCCAAGGG GCTGGGCAAC ATCTCTGCCA
 25286967 AGTATCGCTG GTTCGCCGTC TTCTACCTGA TCATCTTCTT CTTCTGATC CCGCTGACGG
 25287027 TGTCTGGCCT CTCGCTGGCC GGCTGGCGGG TCGTGGTTGG TGTCGGGGTT CCCGTCGTCT
 25287087 TCATCATCAT CCTGGTACTG TGCCTCCGAC TCCTGCAGTC TCGCTGCCCA CGCGTCTGTC
 25287147 CGAAGAAACT CCAGAACTGG AACTTCTGTC CGCTGTGGAT GCGCTCGCTG AAGCCCTGGG
 25287207 ATGCCGTCGT CTCCAAGTTC ACCG

Sequence 35: SD3 (chromosome 11: 57170977 - 57171240)

57170977 CGCGAGAGGA GGGGAATGGC AGGTCCCCGC TTCGCGTACC TTGGAAATAA GCTCATCGTG
 57171037 TTTGGCCAGG TGTGCACGGC AGTGGACACA GCTGTAAGTG CCGTGACAGC GGGGCGAGTA
 57171097 GCTGCGGAAA GTCTTGGTGG GGAGGCAGGC AGGGCCCGGA CCGGGGTCAC AGCTGGGCAT
 57171157 GACGGGCTGG AGGACAATGC CCTGGTGGGC TGGAGGGGCT GGCGCCGTCG AGCCCCCGCA
 57171217 GAGACGGTCG CAGGTGAAGC AGCG

Sequence 36: SD4 (chromosome 11: 1493306 - 1493605)

1493306 CGGTGCTCAC CGGCTCGACC CCCCAGCAAC ACGAGAGACC TCACAGAGGG AGTCACACTA
 1493366 ACCTGGTCCG GGTCCAGAG CGAAACCCCA ACCACTATGC TCACAGCCAG GACCGAGCAG
 1493426 GCTGGGCCAA CGGCAGTCCC TGCCAGCGC CCGGCTCCCT CCGAGTGGCC AGCAGCGCCC
 1493486 TCTGGTGGAG ACTGGCTCGG CCTCCGCGGC ACTGCATTC CACGGCAGTG GTCCATCTAG
 1493546 TCCCCAAGTC CTAGAGGAGG CCCCTCTCTC TCCCTCAGCC CTGGCAGGGT CCTTGGCGCC

Sequence 37: SD5 (chromosome 22: 16491391 - 16491872)

16491391 CTGAGAGCCA TGGCGAGAGC AATGCTAGGC CGGTGAACAG TAGGCTCGAG TTTAGGTTTG
 16491451 AAAGGTGAGG TGAGAGAAAT CGGCAAAAGG AACCCCTGCG CAGATCTCGG GTTCCCTTTAC
 16491511 TTTATAACCG CGGGTTCCGG TTCTCTGCCAG GTGACTGCAC AGTTTCATCCT CATGACCCTT
 16491571 CTCAGCCAAT GGAAGAGAG CGACGCCAG GAAGTCCCGC CCTGTCCCGG CCTGTGGGCG
 16491631 CGTCTCTCGG TCCTTCTACG TCGCTGACTC GTGACCTGAC CCGTATTTTT TTCTAAACT
 16491691 GGGATCTTGG GTAGGAGGAA GAAAAGATAA GGAGTTCTCT TATCTGAAAT TAGTCGGGCT
 16491751 GTTTTGAGGA GTACTGGTTA GGTATATTGG AGAATGTGCC TTTATTGGAA TTTTCTGCAT
 16491811 AATTACACAG GGTATGGGC TTAGGGGTAC ATGATTGTTG GCCGGGCGTG GTGGCTCACG
 16491871 CC

Sequence 38: SD6 (chromosome 4: 155727706 - 155729006)

155727706 TTGCAGGGCT AGCCAGGATA AGCAATGGAT GCTGTTTAAT TCATCTTAGC CTCAGAAATGG
 155727766 ATCAAGAGTA AGTGCTATT CCTCTCAAA GGAAACGCTA ATTAATGAT CCTTGGTGAA
 155727826 ACTCAGGCTG AGGGCTCAGA GGTGTGGTGA TGTAATGGGC TTTGGAATTA GACAGGGACT
 155727886 GAACATTTGC TTTTGAAATT TACTATCTAT GTACCGTTGG AAAATTTACT TAATATCTCT
 155727946 GAATTTTTTT TCTTCAACTG TGGAGTGAGG AAAATAATAC CTACTTTTAG CTAGATGATG
 155728006 GATATAACAC TTTTCTCTGC ATATAGTAGA CACTCAGTGC ATAACATATCG GCTTCTTTT
 155728066 CCTCTACTC AGAAACAAGG ACATCTGGGA CCACAGCCAC ATACTTACCT CCAGTCGTTT
 155728126 CATATCAACC AACTGAGCTC TAACATTTT CTGCAGAAGC TGGATATGCT GACTTTTTTC
 155728186 TATGACTTTG CGCTTCAGGA CTCAATTCT GCTTCTCAGA TCCTCTGACA CTCGGTTGTA
 155728246 GGTATTATCA CGGTCTGAAA TCGAAAATAT GGTATTGAA GTAGCTGCTG AGTGATTTGT

155728306 CTGTAATTGC CAGCAAAAAA GAAAGGAAGA AAGGAAGGAA GGAGAAAGAA AGAAAGAAAG
 155728366 AAAGAAAGAA AGAAAGAAAG AAAGAAAGAA AGAAAGAAAG AGAAAAAGAA AAGAAAGAAAG
 155728426 CTAGCTTGTGTA AATATGCCTA A'TTTTATTTT GGTACAGTT TAATCTGTGA GTTCAAAACC
 155728486 TATGGGGCAT TTGACTTTTG GATAATGTTA TGCCCTGCAG CCTTCCATGA ATGCCAGTTA
 155728546 AGATGTCCTA ATAGCAATTA GTAATCCCAA AGAAATATAG AAGAAGAACT TTCTTTGGAA
 155728606 TTTTAAAGGT GTAATTTGGA GTTAAATAG TTGGTTTGAT TGCATTTCAA TTATTTTATA
 155728666 ACATCCTTAA TCAAGGGACT TGAACATATT GGATTTTCTT ACTGATGAGC TTTTCTTTT
 155728726 AATCTATAGA TTTGAAATGG TTCCCTAAGCT GTTTTGGGTC AACAGGATCA CTCACCTGGCC
 155728786 AGCTAGTGTG GCATCACTGA TTTTAAATGT CAAGTGTGTG TGAAGGTGTA AAAAGGCCAA
 155728846 GCAAACTTGA AGAACTGAG GACTCCCTAG ACTCGCTGTC CATGCCCAGA GTGAATGCAA
 155728906 TGTTCCTTAA CCCTAATGAG TAGACTGTGA GAATGACGTA GCTTGACCCT ATATTTTAAA
 155728966 TTTAAAAATC TACCTAATCA GCTCAGTGGA GTCTGGGAGT T

Sequence 39: L91762 (chromosome 12: 73985133 - 73985792)

73985133 GAACTGAAAT TTCGGAGGGA GCTAAGAGCT CCATGTTTTT ATTTTACTAT TCTTTAAACA
 73985193 AGCAGACACA GTGCTGGTTC CTGATCATGT GGGAGGTGTG TAGGAGAGAA CACAGGTAAA
 73985253 GTAAATATAA AATATAAAGC CATAAATGCC ACAACAGAAA TTCAATGAAC AAACATAGAA
 73985313 GAGAAATACCT TAATTCCCTAG AAGACAGAGA GAAGAGTTTA TGAAAGGCAT CTAGAAGAG
 73985373 GAGATATTTC GACTGATTCT TAAAAGGTGA GAGTACAAGA CTTCCTCTCC AGGAGTCGGC
 73985433 GGAGGGAGGG CAGACTCAA GTGCTCCTGG ACGCAGAGGC TCTTGTGAGA GGGCAGAGAC
 73985493 CGGTTGGAAC AAAGTAAGGA TCTGAGCGAC CCCAACTTTG CAGCCGAGGC CTCCAGCTCC
 73985553 GAGGTGCATA GCAACCTAG GGTTCGGTA GCGTTTCCTC CGTCCGCGAC CCTGGCAGAG
 73985613 CCGCAGAGCC CTCTCCAGA CCGACCACT GCCCAGGCC CCGAGCGACC GGAAGGCTCA
 73985673 GCAGCAGGCC GCGGAGAAGG GCGGCGCCCA CCAGGCCCTT AGCACTGCCC AGTCTGGCCC
 73985733 GGCACAGCTG CTGCAGAAGG CGCACGACGA GCTCGTGTGG TACATGCGGG TCAAGGACCA

Sequence 40: L68346 (chromosome 3: 12175184 - 12175532)

12175184 CGCGGACCTC GGAATCACCA AGTGCCGAGT GGCAGATGTG CTGCTGAGGG TGCGCCGGGG
 12175244 CGCAGCTGCA TGCTCACCC AGCCCCGGGG GCCGCAGCAA CGCCAGCAGC CGCAGCAACA
 12175304 GCACCCAGCT TGGCGCGGGC CGAGGGCTCC CAGGCATGAC ACTGCAGATC CGCAGCTAG
 12175364 CCTGTGAGGT CTGGGGGACT GGACGGCCCC AGCAGGGCTC CTTCCCAAGG CCGTTGTGCC
 12175424 CCTCGCCCCA GGCTTTATGA GGTGGCCCTA AGGGCCAATC CCGCCCCGAC GGGCTCCGCC
 12175484 TTCTTTGGC TCTAGGCCAC CCGCGGAGGC AGGACCCGAG GCTCTTCCG

Sequence 41: L50468 (chromosome 3: 55492699 - 55493120)

55492699 CGCGTGCCAC GGGGAGGAGG ACTGGGGGCG TTTGAGGGGC TCAGCGCACC AGAGGAGTGA
 55492759 GGTGGAGGAG GGCCTTCCCG CGTCTCCTC TTCAATCCAG AGCAGCTCAA CGACGTGGCT
 55492819 CCTTTCTAT GTATCCCTCA AAGCCTTCGC GTCGGATTAA AGGTGTTCTT GATCCTTCTT
 55492879 TACCAACCA GGTGCGGGCC AGCGCTGATG AGGGATGAGG GAGAGGAAAC CTCAGTAGCA
 55492939 AAATTGTTCA GAGGACGTTT GAGGGGCGCG GGGAGCAGCC GGATGCACAC CTAAAGTCTC
 55492999 GCAACACCCA ACTTCTCCTC CGCAGGCAGT CCTTAAGAG AATAGATAAA AAGGCCAAGC
 55493059 AAAGATCCTC TCCCCTCGGC CCCGGGGCAC TGCCACATTC AGTCTAGGCA TCCCCCTCTG
 55493119 CT

Sequence 42: L14432 (chromosome 22: 29866031 - 29866650)

29866031 GCTCGGGTCC GGGGGTGTGG ATGAAGGAGC CCCAGGCAGT CTCATGAGCA CAGAGAGCAC
 29866091 CGTAGGCTGC GGTGAGCTCC GGCTCATCCT CCCAGCTACA TGACTGCAGC CTCTATGCT
 29866151 CATCCCAGCG GGCATGGATC AGGGCCAGTC CCTGAGCATC CTTGGCCAGG AAGCTCAGGT
 29866211 ACCCCAGTGG GTTGCCAGGG ACGGCCTTGG TCAAGTGGCA GGAGGTCTTG TACCAGCGGA
 29866271 GGGCAGGGGA GCCCCCCAGG GCCACCCCA GGAAGCCAG CATCCCAAAC AGCCCTGCCT
 29866331 GAACCCCCAT TCTGCACTGG CCCAGTCCAG TCAGACAAAG CCCCTGGGAT GCCTGCCCTT
 29866391 GGTGGCCCCA CACGGCGGCA GCTGAGCAGT GGAACGGAAG CGGAGCCAG CAGGCCGGT
 29866451 GCGGCGGGAC CAATGAATGG AGCTGCGGGA GGAGGAGGAA AGAGGCTGGA GTATGGGGTG
 29866511 ATCTTGGGCT TGTAAACGAA TCCACAGCC GGGCAGGACG CTGATTGGCT GAGGAGTGCA
 29866571 CTTGCCAGGG CCCACGCCCC CATCTGCCT GCCTTCTCAG CACCATCCAG TCACCTGCTG
 29866631 CCAGCCCTGC CTAGATTGGG

Sequence 43: L4648 (chromosome 1: 35815020 - 35816031)

35815020 CGGCGCGCTG GCTCCCTGGA GCGGGGCGGG ACGCGGCGCG GCGGACTCAC GTGCACAACC
 35815080 GCGCGGGACG GGGCCACGCG GACTCACGTG CACAACCGCG GGACCCAGC GCCAGCGGGA
 35815140 CCCAGCGCGC AGCGGGACCC CAGCGCCAGC GGGACCCAG CGCCAGCGGG ACCCCAGCGC
 35815200 CAGCGGGACC CCAGCGCCAG CGGGACCCCA GCGCCAGCGG GTCTGTGGCC CAGTGGAGCG
 35815260 AGTGGAGCGC TGGCGACCTG AGCGGAGACT GCGCCCTGGA CGCCCCAGC TAGACGTCOA
 35815320 GTTACAGCCC GCGCAGCAGC AGCAAAGGGG AAGGGGCGAG AGCCGGGCAC AGTTGGATCC
 35815380 GGAGGTCTGT ACCCAGGGGA AAGCGTGGGC GGTGACCCA GGGCAGCTGC GCGGCGGAGG
 35815440 CAGGTGGGCT CTTGCTTCCC TGAGGCGGCC CCTCCCCACA CCTGCCCTCG GCGCCCCCAG
 35815500 CAGTTTTCAC CTTGGCCCTC AGCGGAGACT GCGGGATTG GCGTTGCCG CAGCCAGTG
 35815560 GGGAGTGAAT TAGCGCCCTC CTTGCTCCTC GGCCCTTCCG ACGGCACGAG GAACTCCTGT
 35815620 CCTGCCCCAC AGACCTTCGG CTTCCGCCGA GTGCGGTACT GGAGCCTGCC CCGCCAGGGC
 35815680 CCTGGAATCA GAGAAAGTCG CTCTTTGGCC ACCTGAAGCG TCGGATCCCT ACAGTGCCTC
 35815740 CCAGCCTGGG CCGGAGCGGC GGCTGCGTCG CTGAAGGTTG GGTCTCTTGG TCGGAAAGGG

35815800 AGGCAGCTGC AGCCTCAGCC CCACCCAGCA AGCGGCCTTC GCATCGCTGC GGTGGGCGTT
 35815860 CTCGGGCTTC GACTTCGCCA GCGCCGCGGG GCAGAGGCAC CTGGAGCTCG CAGGGCCAG
 35815920 ACCTGGGTTC GAAAAGCTTC GCTGACTGCA GGCAAGCGTC CGGGAGGGGC GGCCAGGCGA
 35815980 AGCCCCGGCG CTTTACCACA CACTTCCGGG TCCCATGCCA GTTGCATCCG CG

Sequence 44: L39664 (chromosome 17: 53710207 - 53710501)

53710207 CGCGCTGAGC GGTGGGTGAG GAGACAGGGG TCATCGTGCA GGTGTGCAAA GGGCAGCAGG
 53710267 GCCCGGCCGT TGTCTTGGA GCGCTGGTTG ACGGCCAGCA GCCCCAGCTG GTTGGACATG
 53710327 TTGCGCAGGT TCCTGGCCAG GGGCTCCTCG CTGCCGTACA CCATGTGTC GTCCACGAAG
 53710387 GAAGTGAGCG CGTTGATCTG GTTGCAGATG GTGATGTGTC TCCCGGGGCA AGCCGGGCGAG
 53710447 GAGCGGAAGA ACGGGATGCA GTCCGGCTTG TTCTTGATGC GGGGGTCATT GGGCG

Sequence 45: L30139 (chromosome 17: 36996273 - 36997073)

36996273 TGCCACACAG AAGCCCATCA CCACCAGCAA AGCCACCACC AAAGCCACCA CCAAGCCAG
 36996333 CACCAAGGCC ACCACCATAT CCTCCCCCAA AGCCACTACC AAAGCTGCTG CTGCTGCTGC
 36996393 TGAAGCCACC GCCATAGCCG CCCCCAGCC CGCAGGCTCC CCCAGAGGAG AAGCGGGAGG
 36996453 ATGAGACAGA CAGGCCGCCC CCGTAGGTGC TGGGGGCGCG GCAGGACCCT CCGGCCAGGA
 36996513 CGGAGGAGAT GCGGCTGGAG CCGCCCCCGA TGCCGCCGCC GATGCCGCG GAGCCCTTCA
 36996573 TGGAGCTGGA GGAGGTGAAC TGGCGGCTGC AGGTGGTCAT GGTGCAGAGG AGGGAGGTGA
 36996633 GCGAGCGAGC AGTTGGCTGA GTGAAGAGAA GGTGCTCGG TAAATTGGAA AGGGATGCGA
 36996693 GTGCTTTATA CTCGTGGGTA GGGGGCGGCT CTGGCACTTT CCATTCCCCT TGGCTTTCAT
 36996753 CACCCACAGG CTAGCGCCAA CTCCAGCCA GGTCCCTCCT CTCATCCGCC TCATCATGTC
 36996813 TGTCAATTTT TACTGGAAAC TCATTGTTTG GAGTGTTTTG GGCTTTCTTG TCCCGCCAGG
 36996873 CGTGATTAC AGGGGGAGGT GTGGGCTGCG AGGCTACACT TTCCCATCGG ACCCTGGGAG
 36996933 TCCAGCCCT CAGGAACCCG CGCACTGGGC TTAGCCAGGG TGACAGAGAG CAGGGCCTCT
 36996993 GCACCTTAAA CCTAGTTACC TGCTAGCTCT CCATGAAGTG AATGGGCTT TAACATCCAC
 36997053 GTAGAGGAAG CCGCTGCTG C

Sequence 46: L55429 (chromosome 5: 1547854 - 1548287)

1547854 CGATGGAGGA CATCGTCATG GTCACAGGGA TGGCGTCGAA GTAGGACGAG TGAGGCGCGA
 1547914 GCGTGAGGAT GGCGCCTCG GTGGGCAGCG CCTGCCGCC CTTCACGGCC ACCCGGTGGA
 1547974 AGCCGCGGCG GAACCATG GTGCGCATGA TGGCTTCAG CAGGAAGTCC ACAACCTGC
 1548034 AAAAGAGGGC GCTGCGTAC GCGGGCACAC GTCCGCAGTC TCGGAGTCTG TGTGAGGCAC
 1548094 AGGGGCGGTC CCACGGGAGA GCCCTCCAGG GCGCAGTCCA GGCCACGGGC TTCCTGTGGT
 1548154 CGCGCGCGCT GGGACATCTG CTTGAGGGAA GAAAAGACGC CGCGCCTTCC TGGCTCCGC
 1548214 CTGTGTCTC GCTGGCTGCG CTCTCACCTA CGAAGGAGGC CGCGGGGCC TGTGTGGTGG
 1548274 TCACGGGCCA GCGC

Sequence 47: L62086 (chromosome 19: 40478392 - 40478802)

40478392 CGTCCATCTC GGCTTCGAA GGCACGTGCG TCTCCATCCC CTGCCGCTTT GACTTCCCGG
 40478452 ATGAGCTGCG GCCCGCTGTG GTGCATGGTG TCTGGTACTT CAATAGCCCC TACCCCAAGA
 40478512 ACTACCCCC GGTGGTCTTC AAGTCGCGCA CCCAAGTAGT CCACAGAGAG TTCCAGGGCC
 40478572 GCAGCGCCTT CCTGGGGGAC CTGGGCTGCG GAAACTGCAC CCTCCTGCTC AGCAACGTCA
 40478632 GCGCCGAGGT GCGCGGAAG TACTACTTCC GTGGGACCT GGGCGGCTAC AACCAGTACA
 40478692 CCTTCTCAGA GCACAGCGTC CTGGATATCG TCAGTGAGTC CCCAGCGGTT GTGCAGGCAC
 40478752 CGGAGCTGG GGCAGCGGG CGGAAGGAG TGTGGCCGA AGGCCTCCCC G

Sequence 48: L76138 (chromosome 19: 3129742 - 3130341)

3129742 CGGGTGGCTC CACCCTGCGT CGGGCCTCAG TCAGCCCCG GGGGAGGCCA TGAACGCCAC
 3129802 GGGGACCCCG GTGGCCCCCG AGTCCTGCCA ACAGCTGGCG GCCGGCGGGC ACAGCCGGCT
 3129862 CATGTGTTCT CACTACAACC ACTCGGGCCG GCTGGCCGGG CGCGGGGGGC CGGAGGATGG
 3129922 CGGCCTGGGG GCCCTGCGGG GGCTGTGCGT GGCCGCCAGC TGCTTGGTGG TGCTGGAGAA
 3129982 CTTGCTGGTG CTGGCGGCA TCACAGCCA CATGCGGTCG CGACGCTGGG TCTACTATTG
 3130042 CCTGGTGAAC ATCAGCTGA GTGACCTGCT CACGGGCGCG GCCTACCTGG CCAACGTGCT
 3130102 GCTGTGCGGG GCGCGCACCT TCCGTCTGGC GCGCGCCAG TGGTTCTTAC GGGAGGGCCT
 3130162 GCTCTTACC GCCCTGGCCG CCTCCACCTT CAGCCTGCTC TTCCTGTCAG GGGAGCGCTT
 3130222 TGCCACCATG GTGCGGCCGG TGGCCGAGAG CGGGGCCACC AAGACCAGCC GCGTCTACGG
 3130282 CTTATCGGC CTCTGCTGGC TGCTGGCCGC GCTGCTGGG ATGCTGCCTT TGCTGGGCTG

Sequence 49: L15952 (chromosome 7: 2740971 - 2741510)

2740971 CGGGGCGAGC ACGTGCACCT CGTGACCACC GGCTCTCTAG AGCCCCAGGC AGAGGCCAG
 2741031 CTCGTCAGTA GGAAGGCAT CGGGAGTGCC AGGGTGAACG TACCCCAAGG GCCCGGCACG
 2741091 ACTGACCTCT CGTGCCTGCT CTCTCCCTTC CTCGCCAAGC CCCCCTGATG TGGGAAGCCA
 2741151 GCGTGAGGCC GGTGGGGCAG CCGCTTCCC GTGGCTGTGC CAAGTCCCC CGGTCTCTCTG
 2741211 CACATCATGC CTCCTTCCAC ACCCTGACAG GAAGCAGCTG GGAGAAGCCG TTGGGTGCAC
 2741271 TCACCTCCCTG ATTTACGACA AGTTCCTTCC TCAGCGCCTC TCTCTCTCTG CTCCTCTGC
 2741331 TCTCTGCCCC TCCCCTGGGC CTCGGGAGGT GCCACGCAAG CCCAAGAAGC ATCAGCATAC
 2741391 TGTCCCTCCC TCTCTGTGG CCACGGGCTC CCCCAGGAG CTGAGAGTAG CAGCAGCTCA
 2741451 CAGCCCAAGC CACCCTTGCC CGTTTCTAGG CAGGTGGTGG CACCAGGCAC GAAGGAAGCA

Sequence 50: L36599 (chromosome 19: 4867561 - 4867784)

```

4867561 CGGTGACCCG GCCGGGTCGA AGGGCAGAGT TCCGCTGTCA CTAGCCCTCC ACCCGTCCTG
4867621 TGTGCTGGGA TGCCCTCGCG GCGCCGTCCA CGCCACCGCC GCGCCCTCTT GTGGGTTCTG
4867681 TCTCCTCCGT GTCTAGGATC CTCCTGCATC CGTTTTTCCT TCCTCCCTTC TCTCCCTCCG
4867741 TCTGTCTTGC CCGCACCTGA GGTGTGCGCA GAGGCGCTGA GACG

```

Sequence 51: L26688 (chromosome 17: 77844647 - 77844962)

```

77844647 CGCGGGCTGC GGGGCTGCGG GGAGGGAGGG GCGGGTTCAG GGCTGTGGGC CCCGCCCCGG
77844707 GCACCGCCTG GATCAGGCCT GAGGCCGGCT GCAGAGAACT TGGGCACCGC GGACCAGATT
77844767 GCTGGTCACC AGGATGAAAA ATAAAAAAGA CCCAAACAGC TCCCCCGACC CCGCCTTCGC
77844827 GCAGGCCCTT GCGTTCCTGC TCAGGGGTTT TTGTTTAAATG AAGAACGTCA AACATCTGGC
77844887 AAGGTCGGAA TGATTTTTCG ATGAGCACCG CCTGGATTCT GCGGTGAAAG CGACTGTGTG
77844947 CGCGCTCAGC GCCTCG

```

Sequence 52: L81528 (chromosome 19: 47395445 - 47395816)

```

47395445 CGGGCGCCGC CCCCTTCCTT CCTCCATCAG CAACAGGCGG CGCCGGCCAG CCTCATAGTC
47395505 AGCCTCATCC ACACTGACCA GCAGGCGAAC AGCCTCCCGG CCCACAGCCT CTCGAGGGC
47395565 CTCAGTCAGG AACACGCCCC GCAGGGCCTG CAGCAGGGCG CCACTCAGGT AGTCGCCCCA
47395625 GAAGGCGTCC AGATAGGAGA GCTCTGAGAA CTTGATGTCA CAAACCACAG AGCCCAGGTC
47395685 CCTTGAGCGC AGCACTGCGG TGGCCTGCCC AAACACGTCC AGCTGCCGCG CCAGCGCCTG
47395745 GGGCCGCCCG GATGCCACGC CCTGCTCCAA GGCTGGCCCA TGCTCGCAGT ACTCTGCTCG
47395805 AACCCGGAGC CG

```

Sequence 53: L36556 (chromosome 19: 50962003 - 50962260)

```

50962003 CGCCCAGGGC CAGGCCGTTG ATGATGACGG GGCCCCCGTT GAGGAGCACT GCTGGGGAGC
50962063 CGCTGGCTGC CAGGAAGCTC CCGTTCACCA GGATGGAGGA GGAAGCCGGG CAAGGCGCGG
50962123 GAGGGCCGGT CCCTGCCAGG AATATGGAGC CCTGGGCAGC GGCCTCGGCG GACACTGGGG
50962183 CCGCCCTCTT CTCCAGGTCC TCAGGACTTC GGCTGGACTC GTCCTCAGTC GTGGGATTCC
50962243 CATCAGACTC GCTGGCCG

```

Sequence 54: L16264 (chromosome 2: 2733750 - 2734147)

```

2733750 CGCCTTTTCC TTCGGTCACT CAGACCGCGC CCGAGGTCCC TGGTGCCCCC GCGGGGGAGC
2733810 TGAGGTTTGC GGTCCCTCCT GGGGTTTCGCT TCCCGCGGGG CTAAACCGCG GCGACCAGGG
2733870 CCCCTTTCTC CACTGGTGCC TTTTCCGGGA AACGCTGCTC CTTAGATGGA CGAATACGTA
2733930 CTCGGTACCC AGCACATCCT GGACGAGTTA ACTTCCTTAA TTTCCATATT TGCCGGAGAA
2733990 CCAGCTGGTT CAGAGCGCAC AGCAAACGGG AGAAGTTAAA CCAGGTTCCTG CGACCCAGAG
2734050 CCCAGGGGTG GCCCCGGGGA CACCCACCTG ACTCCGCACC CCCCACGAGA GGGGAGGATC
2734110 CTTGCAGACC TCACCTTTGC TGGCAACGCT GCGGCCCC

```