

(19) World Intellectual Property  
Organization  
International Bureau



(10) International Publication Number  
**WO 2017/006108 A9**

(43) International Publication Date  
12 January 2017 (12.01.2017)

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

(48) Date of publication of this corrected version:  
26 October 2017 (26.10.2017)

(21) International Application Number:  
PCT/GB2016/052026

(15) Information about Correction:  
see Notice of 26 October 2017 (26.10.2017)

(22) International Filing Date:  
05 July 2016 (05.07.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
62/189,063 06 July 2015 (06.07.2015) US

(71) Applicant: **ILLUMINA CAMBRIDGE LIMITED**  
[GB/GB]; Little Chesterford, Nr Saffron Walden Essex  
CB10 1XL (GB).

(72) Inventor: **FRASER, Louise**; c/o Illumina Cambridge Lim-  
ited, Little Chesterford, Nr Saffron Walden Essex CB10  
1XL (GB).

(74) Agent: **MURGITROYD & COMPANY**; Scotland House,  
165-169 Scotland Street, Glasgow Strathclyde G5 8PL  
(GB).

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

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

Published:  
— with international search report (Art. 21(3))

(54) Title: SAMPLE PREPARATION FOR NUCLEIC ACID AMPLIFICATION

(57) Abstract: Presented herein are methods and compositions for targeted amplification of DNA and sample identification. The methods are particularly useful in validation and quality control of samples and to confirm that WGS sequence data is properly paired with a patient sample prior to delivering sequence data to a physician or to a patient.



## **SAMPLE PREPARATION FOR NUCLEIC ACID AMPLIFICATION**

### **BACKGROUND**

Whole genome sequencing (WGS) is one approach that has application in personalized medicine for identifying disease risks for specific patients, as well as their potential responsiveness to certain therapies (e.g., drug therapy). In a clinical setting, various factors such as complex sample preparation workflows or outsourcing of sample preparation and/or sequencing makes sample mix-ups possible and difficult to detect. One essential condition of a WGS-based approach in clinical applications (e.g., diagnostics, prognostics, and/or therapeutics) is that the sequence information used be unequivocally derived from the specific patient, i.e., ensuring sample identity is essential to allow the accurate assignment of clinical details to sequence data. There is a need for methods of confirming sample identity in WGS data used in clinical applications. There is also a need for sample preparation methods that do not require purification of DNA prior to amplification. Embodiments of the invention set forth herein satisfy these needs and provide other advantages as well.

### **BRIEF SUMMARY**

Presented herein are methods and compositions for targeted amplification of DNA and sample identification. The methods set forth herein are particularly useful, for example, in validation and quality control of samples and to confirm that WGS sequence data is properly paired with a patient sample prior to delivering sequence data to a physician or to a patient. However, it will be appreciated that the methods set forth herein can be used for other suitable applications where quick targeted amplification is desired.

In accordance with the above, in one embodiment presented herein are methods for obtaining nucleic acid sequence information from a biological sample comprising: (a) providing a biological sample comprising different target nucleic acids; (b) contacting the biological sample with a plurality of different probe sets to form hybridization complexes with the different target nucleic acids; (c) amplifying the

nucleic acid from the biological samples to produce amplicons; wherein there is no purification of the nucleic acid from the biological sample prior to the contacting step (b); and (d) obtaining nucleic acid sequence information for a plurality of portions of the amplified sample.

5           In certain aspects, there is no purification of the nucleic acid from the biological sample prior to the amplifying in step (c). In certain aspects, the amplifying in step (c) comprises polymerase chain reaction using at least two amplification primers that are specific for a portion of the sample genome. In certain aspects, the amplifying in step (c) comprises extension and ligation of two probes to form amplification templates.

10           In certain aspects, the supernatant comprising solution-phase hybridization complexes is further contacted with a solid support to form immobilized hybridization complexes when contacted with the extension enzyme and the nucleotides. In certain aspects, the solid support comprises beads. In certain aspects, the solid support comprises a filter plate.

15           In certain aspects, each of the different target nucleic acids comprises, from 3' to 5': contiguous first, second, and third target domains and each probe set comprises: (i) a first probe comprising, from 5' to 3': a first priming sequence and a sequence that is substantially complementary to a first target domain; and (ii) a second probe comprising 5' to 3': a sequence substantially complementary to a third target domain, and a second  
20           priming sequence.

          In certain aspects, the method comprises, prior to step (c), a step of collecting a supernatant comprising solution-phase hybridization complexes from the biological sample. In certain aspects, obtaining nucleic acid sequence information comprises massively parallel sequencing. In certain aspects, obtaining nucleic acid sequence  
25           information comprises detecting the amplicons on the surface of a nucleic acid array.

          In certain aspects, the plurality of probe set comprises at least at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 different probe sets. In certain aspects, the plurality of probe set  
30           comprises a plurality of probes configured to selectively hybridize to polymorphic

regions that are informative of the identity of the source of the sample. In certain aspects, the plurality of probe set comprises a plurality of probes configured to selectively hybridize to regions that comprise cancer-associated polymorphisms. In certain aspects, the sample is a human sample. In certain aspects, the sample comprises  
5 tumor tissue. In certain aspects, the sample comprises normal tissue. In certain aspects, the sample is a blood sample. In certain aspects, the sample comprises dried blood on a porous solid surface. In certain aspects, the porous solid surface comprises filter paper.

Also presented herein is a method of obtaining nucleic acid sequence information from a FFPE sample comprising: (a) providing a FFPE sample comprising  
10 different target nucleic acids embedded within a preserved tissue; (b) contacting the FFPE sample with a plurality of different probe sets to form hybridization complexes with the different target nucleic acids; (c) amplifying the nucleic acid from the FFPE samples to produce amplicons; wherein there is no purification of the nucleic acid from the FFPE sample prior to the contacting step (b); and (d) obtaining nucleic acid  
15 sequence information for a plurality of the amplicons. In particular embodiments, there is no purification of the nucleic acid from the FFPE sample prior to the amplifying in step (c).

In certain aspects, the amplifying in step (c) comprises polymerase chain reaction using at least two amplification primers that are specific for a portion of the  
20 sample genome. In certain aspects, the amplifying in step (c) comprises extension and ligation of two probes to form amplification templates. In certain aspects, the supernatant comprising solution-phase hybridization complexes is further contacted with a solid support to form immobilized hybridization complexes when contacted with the extension enzyme and the nucleotides.

25 In certain aspects, the solid support comprises beads. In certain aspects, the solid support comprises a filter plate. In certain aspects, each of the different target nucleic acids comprises, from 3' to 5': contiguous first, second, and third target domains and each probe set comprises: (i) a first probe comprising, from 5' to 3': a first priming sequence and a sequence that is substantially complementary to a first target domain;  
30 and (ii) a second probe comprising 5' to 3': a sequence substantially complementary to a

third target domain, and a second priming sequence. In certain aspects, the method comprises, prior to step (c), a step of collecting a supernatant comprising solution-phase hybridization complexes from the FFPE sample.

In certain aspects, obtaining nucleic acid sequence information comprises  
5 massively parallel sequencing. In certain aspects, obtaining nucleic acid sequence information comprises detecting the amplicons on the surface of a nucleic acid array.

In certain aspects, the plurality of probe set comprises at least at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80,  
10 90, or at least 100 different probe sets. In certain aspects, the plurality of probe set comprises a plurality of probes configured to selectively hybridize to polymorphic regions that are informative of the identity of the source of the sample. In certain aspects, the plurality of probe set comprises a plurality of probes configured to selectively hybridize to regions that comprise cancer-associated polymorphisms. In  
15 certain aspects, the sample is a human sample. In certain aspects, the sample comprises tumor tissue.

Also presented herein are methods for amplification of nucleic acid from a FFPE sample comprising: (a) providing a FFPE sample comprising nucleic acid embedded within a preserved tissue, the nucleic acid having, from 3' to 5': contiguous first, second,  
20 and third target domains; (b) contacting the FFPE sample with a plurality of different probe sets to form hybridization complexes with the different target nucleic acids, wherein each probe set comprises: (i) a first probe comprising, from 5' to 3': a first priming sequence and a sequence that is substantially complementary to the first target domain; and (ii) a second probe comprising 5' to 3': a sequence substantially  
25 complementary to the third target domain, and a second priming sequence; (c) contacting the hybridization complexes with an extension enzyme and nucleotides, wherein the first probes are extended along the second target domains of hybridization complexes formed in (b); (d) ligating the extended first probes to the second probes to form amplification templates; and (e) amplifying the amplification templates with first  
30 and second primers that are complementary to the first priming sequence and the second

priming sequence to produce amplicons and obtaining nucleic acid sequence information for a plurality of the amplicons.

In certain aspects, the method comprises, prior to step (c), a step of collecting a supernatant comprising solution-phase hybridization complexes from the FFPE sample.

- 5 In certain aspects, the supernatant comprising solution-phase hybridization complexes is further contacted with a solid support to form immobilized hybridization complexes when contacted with the extension enzyme and the nucleotides. In certain aspects, the solid support comprises beads. In certain aspects, the solid support comprises a filter plate. In certain aspects, obtaining nucleic acid sequence information comprises
- 10 massively parallel sequencing. In certain aspects, obtaining nucleic acid sequence information comprises detecting the amplicons on the surface of a nucleic acid array.

- In certain aspects, the plurality of probe set comprises at least at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80,
- 15 90, or at least 100 different probe sets. In certain aspects, the plurality of probe set comprises a plurality of probes configured to selectively hybridize to polymorphic regions that are informative of the identity of the source of the sample. In certain aspects, the plurality of probe set comprises a plurality of probes configured to selectively hybridize to regions that comprise cancer-associated polymorphisms. In
- 20 certain aspects, the sample is a human sample. In certain aspects, the sample comprises tumor tissue. In certain aspects, the sample comprises normal tissue.

- Also presented herein is a method for nucleic acid sample identification comprising: (a) providing a nucleic acid-containing cellular sample; (b) lysing cells of the sample with a lysis reagent to liberate nucleic acid from within the cells of the
- 25 cellular sample, thereby forming a lysate; (c) amplifying the nucleic acid from the lysed samples; wherein there is no purification of the nucleic acid from the lysate prior to beginning the amplification step (c); and (d) obtaining nucleic acid sequence information for a plurality of portions of the amplified sample, and comparing the sequence information to a second set of sequence information.

In certain aspects, the nucleic acid is DNA. In certain aspects, the sample is a blood sample. In certain aspects, the sample comprises dried blood. In certain aspects, the sample comprises a FFPE tissue sample. In certain aspects, the second set of sequence information comprises a whole genome sequence. In certain aspects, the second set of sequence information comprises exome sequence information.

In certain aspects, the amplifying comprises a targeted amplification reaction. In certain aspects, the targeted amplification reaction comprises extension and ligation of two probes. In certain aspects, the targeted amplification reaction comprises polymerase chain reaction using at least two amplification primers that are specific for a portion of the sample genome.

Also presented herein is a method of tracking the identity of a biological sample during different stages of sample processing, comprising: (a) providing a nucleic acid-containing cellular sample; (b) separating a portion of the sample into a first portion and a second portion and obtaining a first set of nucleic acid sequence information from the first portion the biological sample according to any of the embodiments presented herein, wherein the first set of nucleic acid sequence information comprises identity informative sequence information; (c) purifying nucleic acid from the second portion and obtaining a second set of sequence information; and (d) using computer-assisted logic, comparing the identity informative sequence information from the first set of nucleic acid sequence information sequence information to the second set of sequence information to confirm that the first and second sets of sequence information were obtained from the same source.

In certain aspects, the identity informative sequence information comprises SNP genotype information for at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 unique SNPs. In certain aspects, the second set of sequence information comprises SNP genotype information for at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 unique SNPs.

In certain aspects, the nucleic acid is DNA. In certain aspects, the sample is a blood sample. In certain aspects, the sample comprises dried blood. In certain aspects, the sample comprises dried blood on a porous solid surface. In certain aspects, the sample comprises a FFPE tissue sample. In certain aspects, the second set of sequence information comprises a whole genome sequence. In certain aspects, the second set of sequence information comprises exome sequence information.

Also presented herein is a method of confirming the source of two different biological samples comprising: (a) providing a first nucleic acid-containing cellular sample; (b) obtaining a first set of nucleic acid sequence information from the first portion the biological sample according to any of the embodiments presented herein, wherein the first set of nucleic acid sequence information comprises identity informative sequence information; (c) providing a second nucleic acid sample comprising purified nucleic acid and obtaining a second set of sequence information; and (d) using computer-assisted logic, comparing the identity informative sequence information from the first set of nucleic acid sequence information sequence information to the second set of sequence information to confirm that the first and second sets of sequence information were obtained from the same individual.

In certain aspects, the identity informative sequence information comprises SNP genotype information for at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 unique SNPs. In certain aspects, the second set of sequence information comprises SNP genotype information for at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 unique SNPs.

In certain aspects, the nucleic acid is DNA. In certain aspects, the sample is a blood sample. In certain aspects, the sample comprises dried blood. In certain aspects, the sample comprises dried blood on a porous solid surface. In certain aspects, the sample comprises a FFPE tissue sample. In certain aspects, the second set of sequence

information comprises a whole genome sequence. In certain aspects, the second set of sequence information comprises exome sequence information.

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

Additional embodiments can be found in U.S. Provisional Patent Application No. 61/189,063 filed on July 6, 2015, the contents of which are incorporated herein by reference in their entireties.

10

### **BRIEF DESCRIPTION OF THE DRAWINGS**

Figure 1 illustrates a flow diagram of an example of method of using TruSeq® Custom Amplicon for preparation of targeted amplicon libraries and subsequent sequencing for sample identification.

Figure 2 shows an Agilent TapeStation image of a gel of the PCR amplification products generated from each DBS punch and genomic DNA samples using the method of Figure 1.

Figure 3 shows a screenshot of an example of the MiSeq status pane with quality metrics for the sequence run using the pooled DBS library samples of Figure 2.

Figure 4 shows a plot of the percent reads (passing filter) identified as a function of index number for the sequencing run of Figure 3.

Figure 5 shows a table of the SNP calls for 7 donor samples (i.e., CS219, CS220, CS221, CS222, CS223, CS224, and CS225) of Table 2 for both genomic DNA (gDNA) and DBS sample ("spot").

Figures 6A and 6B show a color key and histogram and a plot of identity by state (IBS) calculations, respectively, for the donor samples shown in the table of Figure 5.

Figures 7A and 7B show a color key and histogram and a plot of identity by state (IBS) calculations, respectively, and show an example of discriminating DBS samples of Table 2 based on SNP data.

Figure 8 shows an Agilent TapeStation image of a gel of the PCR amplification products from FFPE and genomic DNA samples in pool A and a BioAnalyzer image of the PCR amplification products from FFPE and genomic DNA samples in pool B.

Figure 9 illustrates a flow diagram of an example of method of using multiplex  
5 targeted PCR amplification for preparation of targeted amplicon libraries and subsequent sequencing for sample identification.

Figures 10A and 10B show a BioAnalyzer image of a gel of the PCR amplification products generated from gDNA and a BioAnalyzer image of the PCR amplification products generated from each DBS punch, respectively, using the method  
10 of Figure 9.

Figure 11 shows a screenshot of illumina sequence analysis viewer (SAV) software summarizing quality metrics for the MiSeq sequence run using the DBS and genomic amplicon libraries of Figures 10A and 10B.

Figure 12 shows a plot of the percent reads (passing filter) identified as a  
15 function of index number for the sequencing run of Figure 11.

Figures 13A and 13B show a color key and histogram and a plot of identity by state (IBS) calculations, respectively, for the donor samples of Table 2 used to evaluate method 900 of Figure 9.

Figures 14A and 14B show a color key and histogram and a plot of identity by  
20 state (IBS) calculations, respectively, and show an example of discriminating DBS samples of Table 2 based on SNP data.

Figure 15 shows a family tree diagram of the Platinum Genomes family.

Figures 16A and 16B show a color key and histogram and a plot of identity by  
state (IBS) calculations, respectively, and show an example of high resolution genomic  
25 DNA sample discrimination based on SNP data.

### **DETAILED DESCRIPTION**

The present disclosure provides methods for targeted amplification of DNA and  
30 sample identification. The methods set forth herein are particularly useful, for example,

in validation and quality control of samples and to confirm that WGS sequence data is properly paired with a patient sample prior to delivering sequence data to a physician or to a patient. However, it will be appreciated that the methods set forth herein can be used for other suitable applications where quick targeted amplification is desired.

5           In some embodiments, the present disclosure relates to methods for preparing samples for subsequent nucleic acid (e.g. DNA) amplification, which methods are simpler to perform than existing methods. In particular, the present invention relates to methods wherein purifying nucleic acid (e.g. DNA) from a sample is not required prior to amplification.

10           The methods set forth herein are particularly useful, for example, in validation and quality control of samples and to confirm that WGS sequence data is properly paired with a patient sample prior to delivering sequence data to a physician or to a patient. However, it will be appreciated that the methods set forth herein can be used for other suitable applications where quick targeted amplification is desired.

15           Accordingly, in some embodiment presented herein are methods for obtaining nucleic acid sequence information from a biological sample comprising: (a) providing a biological sample comprising different target nucleic acids; (b) contacting the biological sample with a plurality of different probe sets to form hybridization complexes with the different target nucleic acids; (c) amplifying the nucleic acid from the biological  
20           samples to produce amplicons; wherein there is no purification of the nucleic acid from the biological sample prior to the contacting step (b); and (d) obtaining nucleic acid sequence information for a plurality of portions of the amplified sample.

          In some of the above embodiments, there is no purification of the nucleic acid from the biological sample prior to the amplifying in step (c). Instead, amplification  
25           takes place in the presence of the components of the sample, including cellular debris and material used for storage of the biological sample. For example, in some embodiments, the amplification reaction takes place in the presence of formalin and paraffin components, in addition to cellular debris. In certain aspects, the amplifying in step (c) comprises polymerase chain reaction using at least two amplification primers  
30           that are specific for a portion of the sample genome. In certain aspects, the amplifying

in step (c) comprises extension and ligation of two probes to form amplification templates.

In certain aspects, the supernatant comprising solution-phase hybridization complexes is further contacted with a solid support to form immobilized hybridization complexes when contacted with the extension enzyme and the nucleotides. For  
5 example, the solid support can be beads, particles, a filter plate, and the like.

In certain aspects, each of the different target nucleic acids comprises, from 3' to 5': contiguous first, second, and third target domains and each probe set comprises: (i) a first probe comprising, from 5' to 3': a first priming sequence and a sequence that is  
10 substantially complementary to a first target domain; and (ii) a second probe comprising 5' to 3': a sequence substantially complementary to a third target domain, and a second priming sequence.

In certain aspects, the method comprises, prior to the amplifying step, a step of collecting a supernatant comprising solution-phase hybridization complexes from the  
15 biological sample. In certain aspects, obtaining nucleic acid sequence information comprises massively parallel sequencing. In certain aspects, obtaining nucleic acid sequence information comprises detecting the amplicons on the surface of a nucleic acid array.

In some embodiments presented herein, the method is a method of obtaining  
20 nucleic acid sequence information from a FFPE sample comprising: (a) providing a FFPE sample comprising different target nucleic acids embedded within a preserved tissue; (b) contacting the FFPE sample with a plurality of different probe sets to form hybridization complexes with the different target nucleic acids; (c) amplifying the nucleic acid from the FFPE samples to produce amplicons; wherein there is no  
25 purification of the nucleic acid from the FFPE sample prior to the contacting step (b); and (d) obtaining nucleic acid sequence information for a plurality of the amplicons. In particular embodiments, there is no purification of the nucleic acid from the FFPE sample prior to the amplifying in step (c).

In some embodiments are methods for amplification of nucleic acid from a  
30 FFPE sample comprising: (a) providing a FFPE sample comprising nucleic acid

embedded within a preserved tissue, the nucleic acid having, from 3' to 5': contiguous first, second, and third target domains; (b) contacting the FFPE sample with a plurality of different probe sets to form hybridization complexes with the different target nucleic acids, wherein each probe set comprises: (i) a first probe comprising, from 5' to 3': a first priming sequence and a sequence that is substantially complementary to the first target domain; and (ii) a second probe comprising 5' to 3': a sequence substantially complementary to the third target domain, and a second priming sequence; (c) contacting the hybridization complexes with an extension enzyme and nucleotides, wherein the first probes are extended along the second target domains of hybridization complexes formed in (b); (d) ligating the extended first probes to the second probes to form amplification templates; and (e) amplifying the amplification templates with first and second primers that are complementary to the first priming sequence and the second priming sequence to produce amplicons and obtaining nucleic acid sequence information for a plurality of the amplicons.

15 In certain aspects, the method comprises, prior to step (c), a step of collecting a supernatant comprising solution-phase hybridization complexes from the FFPE sample. In certain aspects, the supernatant comprising solution-phase hybridization complexes is further contacted with a solid support to form immobilized hybridization complexes when contacted with the extension enzyme and the nucleotides. In certain aspects, the solid support comprises beads. In certain aspects, the solid support comprises a filter plate. In certain aspects, obtaining nucleic acid sequence information comprises massively parallel sequencing. In certain aspects, obtaining nucleic acid sequence information comprises detecting the amplicons on the surface of a nucleic acid array.

25 In some embodiments presented herein, the method is a method for nucleic acid sample identification comprising: (a) providing a nucleic acid-containing cellular sample; (b) lysing cells of the sample with a lysis reagent to liberate nucleic acid from within the cells of the cellular sample, thereby forming a lysate; (c) amplifying the nucleic acid from the lysed samples; wherein there is no purification of the nucleic acid from the lysate prior to beginning the amplification step (c); and (d) obtaining nucleic

acid sequence information for a plurality of portions of the amplified sample, and comparing the sequence information to a second set of sequence information.

As used herein, “targeted amplification” can refer to any amplification method used to amplify target nucleic acid sequences of interest. In some embodiments, the targeted amplification can include amplicon preparation and sequencing on a solid support, as set forth in any one of the disclosures of PCT/US2014/071263, filed on December 18, 2014, entitled “AMPLICON PREPARATION AND SEQUENCING ON SOLID SUPPORTS” and PCT/EP2014/079145, filed on December 23, 2014, entitled “POLYNUCLEOTIDE MODIFICATION ON SOLID SUPPORTS”, which are incorporated herein by reference in their entireties.

In some embodiments, targeted amplification can comprise hybridization of two or more probes to the same strand of a target nucleic acid sequence of interest, followed by extension of the 3’ end of one of the probes and ligation of the extended probe to the other probe, as set forth generally in the materials of U.S. Patent No. 7,803,537, U.S. Patent No. 8,003,354, and U.S. Patent No. 8,906,626, the disclosures of which are incorporated herein by reference in their entireties.

In some embodiments, nucleic acid preparation and amplification is performed without purification of nucleic acid prior to amplification. Some methods of nucleic acid preparation and amplification which are useful in the methods described herein can be found in the disclosures of PCT/GB2015/051674, filed June 9, 2015, entitled “SAMPLE PREPARATION FOR NUCLEIC ACID AMPLIFICATION” and U.S. Application Ser. No. 62/167,463, filed on May 28, 2015, entitled “SURFACE-BASED TAGMENTATION”, which are incorporated herein by reference in their entireties.

As used herein, a nucleic acid sample can be any sample comprising nucleic acid. The nucleic acid may be, for example, DNA or RNA. In some embodiments, the sample comprises whole cells, lysed cells, cellular components, or any mixture thereof. In some embodiments, the sample is a blood sample. In some embodiments, the sample is a whole blood sample. Suitable samples can be obtained from tissue, bodily fluid, and/or any other nucleic-acid comprising specimen. In some embodiments, the invention provides methods of preparing non-blood samples, such as tissue samples

(e.g. formalin fixed paraffin-embedded (FFPE) samples) for DNA amplification. Such tissue samples may be tumor samples. Other samples may be biopsies, or aspirates, and the like.

The present disclosure provides methods of validating the identity of a DNA sample in genetic testing assay. In one example, the genetic testing assay is a multiplexed whole genome sequencing (WGS) diagnostic assay. In various embodiments, the methods of the invention use sequencing of targeted amplicon libraries to confirm the identity of a sample in a genetic testing assay.

In some embodiments, the targeted amplicon library is prepared using a panel of probes or primers that can be used to selectively amplify a plurality of genomic regions that, alone or in combination, provide information about the identity of the sample. For example, in some embodiments, the panel of probes or primers are used to distinguish the identity of one sample from other samples. In some embodiments, the panel of probes or primers are used to identify a human from which a sample was obtained. In some embodiments, the human can be distinguished from a plurality of other humans. In some embodiments, the human can be uniquely identified. For example, in some embodiments, a panel of probes or primers is used to selectively amplify a plurality of genomic regions that, alone or in combination, uniquely identify the biological source, such as a human donor, from more than 10, 100, 1000,  $10^4$ ,  $10^5$ ,  $10^6$ ,  $10^7$ ,  $10^8$ ,  $10^9$ ,  $10^{10}$ , or more than  $10^{11}$  other human donors. In some embodiments, a panel of probes or primers is used to selectively amplify a plurality of genomic regions that, alone or in combination, uniquely identify the biological source, such as a human donor, from  $10^8$ ,  $10^9$ ,  $10^{10}$ , more than  $10^{11}$  other human donors. As used herein, the term uniquely identify refers to the ability of one or more markers that, alone or in combination, distinguish one particular sample from any other sample.

In some embodiments, the targeted amplicon libraries are prepared using a panel of probes or primers that selectively amplify polymorphic regions. Any polymorphic region may be used in the panels presented herein. For example, in some embodiments, the polymorphic regions are single nucleotide polymorphisms (SNPs). In some embodiments, the polymorphic regions are repeated regions, such as short tandem

repeats (STRs). In some embodiments, the panel selectively amplifies a panel that includes a combination of different types of polymorphic regions, such as, for example, a combination of one or more SNPs and one or more STRs.

Any suitable combination of polymorphic regions, and any suitable number of polymorphic regions that provides a desired level of distinction between individuals can be used. In some embodiments, the targeted amplicon libraries are prepared using a panel of probes or primers for detection of identity-informative single nucleotide polymorphisms (iSNPs), also referred to herein as identity SNPs. Any suitable panel of identity SNPs, as are known in the art, may be used in the methods presented herein, such as, for example, those as described in the disclosures of Kidd KK, et al. Forensic Sci Int. 2006; 164(1): 20-32, and by Sanchez JJ, et al. Electrophoresis 2006; 27(9):1713-1724, the disclosure of each of which is incorporated herein by reference in its entirety. In some embodiments, such as those described in examples presented hereinbelow, a panel such as the 45-plex ID SNP subset from ForenSeq set (Illumina, Inc.) may be used in the methods presented herein.

In some embodiments, the identity SNPs are used to discriminate between samples in a multiplex genetic testing assay. The content of an identity SNP panel and/or the number of SNPs in a panel may be selected for a specific application. Primer design (e.g., primer position) for generation of SNP amplicon libraries may be selected for rapid library sequencing. In some embodiments, the SNP panel comprises at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 unique SNPs for positive identification of the sample from one or more samples. In some embodiments, the SNP panel comprises at least 19 different SNPs. In some embodiments, the SNP panel comprises at least 52 different SNPs. In one example, an identity SNP panel includes 24 different SNPs. In another example, an identity SNP panel includes 45 different SNPs.

In accordance with the above, in some embodiments, the method comprises contacting a nucleic acid sample with a plurality of different primers or probes

configured to selectively amplify a desired panel of polymorphic regions, such as, for example, a combination of one or more SNPs and/or one or more STRs.

By way of a non-limiting example, the panel can comprise one or more of the identity informative SNPs set forth in Table 1 below, in any combination.

5

Table 1 Identity Informative SNPs

| Locus      | Amplicon Length (bp) | Chromosome | Amplicon Start Position | Amplicon End Position | 2800M Control Alleles |
|------------|----------------------|------------|-------------------------|-----------------------|-----------------------|
| rs10495407 | 109                  | 1          | 238439234               | 238439342             | G                     |
| rs1294331  | 85                   | 1          | 233448359               | 233448443             | GA                    |
| rs1413212  | 64                   | 1          | 242806767               | 242806830             | G                     |
| rs1490413  | 98                   | 1          | 4367256                 | 4367353               | A                     |
| rs560681   | 90                   | 1          | 160786641               | 160786730             | AG                    |
| rs891700   | 115                  | 1          | 239881850               | 239881964             | AG                    |
| rs1109037  | 118                  | 2          | 10085691                | 10085808              | G                     |
| rs12997453 | 100                  | 2          | 182413195               | 182413294             | A                     |
| rs876724   | 119                  | 2          | 114945                  | 115063                | C                     |
| rs907100   | 115                  | 2          | 239563542               | 239563656             | CG                    |
| rs993934   | 120                  | 2          | 124109120               | 124109239             | C                     |
| rs1355366  | 119                  | 3          | 190806041               | 190806159             | AG                    |
| rs1357617  | 120                  | 3          | 961696                  | 961815                | AT                    |
| rs2399332  | 157                  | 3          | 110300999               | 110301155             | AC                    |

| Locus      | Amplicon Length (bp) | Chromosome | Amplicon Start Position | Amplicon End Position | 2800M Control Alleles |
|------------|----------------------|------------|-------------------------|-----------------------|-----------------------|
| rs4364285  | 98                   | 3          | 32417576                | 32417673              | G                     |
| rs6444724  | 120                  | 3          | 193207306               | 193207425             | T                     |
| rs1979255  | 102                  | 4          | 190318007               | 190318108             | G                     |
| rs2046361  | 120                  | 4          | 10968994                | 10969113              | A                     |
| rs279844   | 167                  | 4          | 46329584                | 46329750              | AT                    |
| rs6811238  | 120                  | 4          | 169663541               | 169663660             | G                     |
| rs13182883 | 169                  | 5          | 136633282               | 136633420             | AG                    |
| rs139606   | 104                  | 5          | 17374845                | 17374948              | A                     |
| rs251934   | 97                   | 5          | 174778619               | 174778715             | T                     |
| rs338882   | 157                  | 5          | 178690899               | 178690755             | C                     |
| rs717302   | 110                  | 5          | 2879333                 | 2879442               | G                     |
| rs13218440 | 170                  | 6          | 12059928                | 12060097              | AG                    |
| rs1336071  | 120                  | 6          | 94537182                | 94537301              | G                     |
| rs214955   | 120                  | 6          | 152697629               | 152697748             | G                     |
| rs727811   | 115                  | 6          | 165045254               | 165045368             | A                     |
| rs321198   | 165                  | 7          | 137029715               | 137029879             | T                     |
| rs6953448  | 120                  | 7          | 4310285                 | 4310404               | CT                    |
| rs737681   | 120                  | 7          | 155990742               | 155990861             | T                     |
| rs917118   | 109                  | 7          | 4456953                 | 4457061               | C                     |
| rs10092491 | 116                  | 8          | 29411037                | 29411152              | CT                    |
| rs2056277  | 104                  | 8          | 139399038               | 139399141             | C                     |
| rs4606077  | 151                  | 8          | 144656710               | 144656860             | CT                    |
| rs763869   | 85                   | 8          | 1375576                 | 1375660               | CT                    |
| rs1015250  | 117                  | 9          | 1823702                 | 1823818               | G                     |
| rs10776839 | 103                  | 9          | 137417271               | 137417373             | G                     |
| rs1360288  | 119                  | 9          | 128967994               | 128968112             | C                     |
| rs1463729  | 99                   | 9          | 126881396               | 126881494             | GA                    |
| rs7041158  | 115                  | 9          | 27985907                | 27986021              | C                     |
| rs3780962  | 94                   | 10         | 17193284                | 17193377              | T                     |
| rs735155   | 170                  | 10         | 3374133                 | 3374302               | A                     |
| rs740598   | 120                  | 10         | 118506839               | 118506958             | AG                    |
| rs826472   | 153                  | 10         | 2406511                 | 2406663               | T                     |
| rs964681   | 105                  | 10         | 132698394               | 132698498             | CT                    |
| rs10488710 | 118                  | 11         | 115207134               | 115207251             | CG                    |
| rs1498553  | 111                  | 11         | 5708981                 | 5709091               | CT                    |
| rs2076848  | 118                  | 11         | 134667502               | 134667619             | AT                    |
| rs901398   | 90                   | 11         | 11096173                | 11096282              | T                     |
| rs10773760 | 99                   | 12         | 130761623               | 130761721             | AG                    |
| rs2107612  | 103                  | 12         | 888262                  | 888364                | AG                    |
| rs2111980  | 94                   | 12         | 106328186               | 106328279             | G                     |
| rs2269355  | 65                   | 12         | 6945881                 | 6945945               | C                     |
| rs2920816  | 157                  | 12         | 40862976                | 40863132              | T                     |
| rs1058083  | 76                   | 13         | 100038193               | 100038268             | AG                    |
| rs1335873  | 109                  | 13         | 20901665                | 20901773              | T                     |
| rs1886510  | 116                  | 13         | 22374646                | 22374761              | CT                    |
| rs354439   | 170                  | 13         | 106938320               | 106938489             | T                     |
| rs1454361  | 118                  | 14         | 25850765                | 25850882              | AT                    |

| Locus     | Amplicon Length (bp) | Chromosome | Amplicon Start Position | Amplicon End Position | 2800M Control Alleles |
|-----------|----------------------|------------|-------------------------|-----------------------|-----------------------|
| rs4530059 | 170                  | 14         | 104769099               | 104769268             | G                     |
| rs722290  | 101                  | 14         | 53216686                | 53216786              | G                     |
| rs873196  | 114                  | 14         | 98845506                | 98845619              | CT                    |
| rs1528460 | 115                  | 15         | 55210664                | 55210778              | T                     |
| rs1821380 | 118                  | 15         | 39313343                | 39313460              | G                     |
| rs8037429 | 63                   | 15         | 53616876                | 53616938              | T                     |
| rs1382387 | 89                   | 16         | 80106318                | 80106406              | CT                    |
| rs2342747 | 104                  | 16         | 5868645                 | 5868748               | AG                    |
| rs430046  | 119                  | 16         | 78016980                | 78017098              | C                     |
| rs729172  | 104                  | 16         | 5606153                 | 5606256               | C                     |
| rs740910  | 113                  | 17         | 5706552                 | 5706664               | A                     |
| rs8078417 | 143                  | 17         | 80461847                | 80461989              | CT                    |
| rs938283  | 98                   | 17         | 77468433                | 77468530              | T                     |
| rs9905977 | 170                  | 17         | 2919324                 | 2919493               | G                     |
| rs1024116 | 98                   | 18         | 75432317                | 75432414              | A                     |
| rs1493232 | 75                   | 18         | 1127945                 | 1128019               | A                     |
| rs1736442 | 153                  | 18         | 55225698                | 55225850              | G                     |
| rs9951171 | 119                  | 18         | 9749789                 | 9749907               | G                     |
| rs576261  | 76                   | 19         | 39559780                | 39559855              | AC                    |
| rs719366  | 170                  | 19         | 28463281                | 28463450              | T                     |
| rs1005533 | 158                  | 20         | 39487066                | 39487223              | A                     |
| rs1031825 | 126                  | 20         | 4447416                 | 4447541               | C                     |
| rs1523537 | 117                  | 20         | 51296076                | 51296192              | C                     |
| rs445251  | 119                  | 20         | 15124865                | 15124983              | CG                    |
| rs221956  | 97                   | 21         | 43606933                | 43607029              | C                     |
| rs2830795 | 114                  | 21         | 28608089                | 28608202              | A                     |
| rs2831700 | 79                   | 21         | 29679639                | 29679717              | A                     |
| rs722098  | 101                  | 21         | 16685561                | 16685661              | AG                    |
| rs914165  | 156                  | 21         | 42415865                | 42416020              | AG                    |
| rs1028528 | 78                   | 22         | 48362256                | 48362333              | AC                    |
| rs2040411 | 68                   | 22         | 47836378                | 47836445              | A                     |
| rs733164  | 120                  | 22         | 27816711                | 27816830              | AG                    |
| rs987640  | 120                  | 22         | 33559450                | 33559569              | AT                    |

A sample in a genetic testing assay may be any sample that comprises nucleic acid, such as a biological sample. Non-limiting examples of biological samples can include whole organisms as well as a sample obtained from a patient. The biological sample can be obtained from any biological fluid or tissue and can be in a variety of forms, including liquid fluid and tissue, solid tissue, and preserved forms such as dried, frozen, and fixed forms. The sample may be of any biological tissue, cells or fluid.

Such samples include, but are not limited to, sputum, blood, serum, plasma, blood cells (e.g., white cells), ascitic fluid, urine, saliva, tears, sputum, vaginal fluid (discharge), washings obtained during a medical procedure (e.g., pelvic or other washings obtained during biopsy, endoscopy or surgery), tissue, nipple aspirate, core or fine needle biopsy samples, cell-containing body fluids, free floating nucleic acids, peritoneal fluid, and pleural fluid, or cells therefrom. Biological samples may also include sections of tissues such as frozen or fixed sections taken for histological purposes or microdissected cells or extracellular parts thereof. In some embodiments, the sample can be a blood sample, such as, for example, a whole blood sample. In another example, the sample is an unprocessed dried blood spot (DBS) sample. In yet another example, the sample is a formalin-fixed paraffin-embedded (FFPE) sample. In yet another example, the sample is a saliva sample. In yet another example, the sample is a dried saliva spot (DSS) sample.

In some embodiments, a panel of identity markers is included in a panel designed to screen for a disease, such as, for example a somatic cancer mutation. As a result, the results of sequencing the combined panels not only provides information about the disease state, but also provides information about the identity of the individual from which the sample was obtained. The combination of disease-informative sequence information and identity-informative sequence information allows the sequence information about the disease to be tied to the identity of the individual. This provides significant advantages in storing and management of sequence data from individuals because the identity informative sequence information is a component of the sequence information obtained when screening for the disease state. Additionally, the identity informative sequence information can be advantageous when comparing the results of the disease screen with other sequence information, such as whole genome sequence information, exome sequence information, or a panel of sequence information that is directed to another disease screen.

In some embodiments, the methods and compositions presented herein allow for tracking the identity of a biological sample during different stages of sample processing. For example, a method of tracking can comprise: (a) providing a nucleic acid-

containing cellular sample; (b) separating a portion of the sample into a first portion and a second portion and obtaining a first set of nucleic acid sequence information from the first portion the biological sample according to any of the embodiments presented herein, wherein the first set of nucleic acid sequence information comprises identity  
5 informative sequence information; (c) purifying nucleic acid from the second portion and obtaining a second set of sequence information; and (d) using computer-assisted logic, comparing the identity informative sequence information from the first set of nucleic acid sequence information sequence information to the second set of sequence information to confirm that the first and second sets of sequence information were  
10 obtained from the same source.

In some embodiments, it may be desired to confirm the identity of the source of two different samples. For example, one method can comprise: (a) providing a first nucleic acid-containing cellular sample; (b) obtaining a first set of nucleic acid sequence information from the first portion the biological sample according to any of the  
15 embodiments presented herein, wherein the first set of nucleic acid sequence information comprises identity informative sequence information; (c) providing a second nucleic acid sample comprising purified nucleic acid and obtaining a second set of sequence information; and (d) using computer-assisted logic, comparing the identity informative sequence information from the first set of nucleic acid sequence  
20 information sequence information to the second set of sequence information to confirm that the first and second sets of sequence information were obtained from the same individual.

#### Selectively Amplifying

25 In embodiments presented herein, a panel of target nucleic acids is selectively amplified from a sample. In some embodiments, selectively amplifying can include one or more non-selective amplification steps. For example, an amplification process using random or degenerate primers can be followed by one or more cycles of amplification using target-specific primers.

In one embodiment, the methods of the invention use TruSeq® Custom Amplicon (TSCA) (Illumina, Inc.) for preparation of targeted amplicon libraries for subsequent sequencing and sample validation. The TSCA library preparation protocol uses target specific oligonucleotide probes and a ligation/ extension reaction to generate targeted amplicons. For example, in some embodiments, the method can comprise (a) providing a sample having the different target nucleic acids comprising DNA, wherein each nucleic acid comprises, from 3' to 5': contiguous first, second, and third target domains; (b) contacting the sample with a plurality of at least 100 different probe sets to form hybridization complexes with the different target nucleic acids, wherein each probe set comprises: (i) a first probe comprising, from 5' to 3': a first priming sequence and a sequence that is substantially complementary to the first target domain; and (ii) a second probe comprising 5' to 3': a sequence substantially complementary to the third target domain, and a second priming sequence, wherein at least one probe in each of the different probe sets contains a distinct adapter sequence not native to the target nucleic acid; (c) contacting the hybridization complexes with an extension enzyme and nucleotides, wherein the first probes are extended along the second target domains of hybridization complexes formed in (b), wherein the hybridization complexes are immobilized on a solid support when contacted with the extension enzyme and the nucleotides; (d) ligating the extended first probes to the second probes to form amplification templates; (e) amplifying the amplification templates with first and second primers that are complementary to the first priming sequence and the second priming sequence to produce amplicons; and (f) detecting the amplicons on the surface of a nucleic acid array that is different from the solid support that is immobilized to the hybridization complexes.

In another embodiment, the methods of the invention use targeted multiplex PCR amplification for preparation of amplicon libraries for subsequent sequencing and sample validation. Although TSCA and targeted multiplex PCR are described herein as exemplary embodiments, it will be appreciated by one of skill in the art that any of a number of known methodologies for selective capture and/or selective amplification can

be used in the methods presented herein for preparation of an amplicon library for subsequent sequencing and sample validation.

As used herein, the terms “amplifying,” “amplify,” “amplification” and like terms refer to producing one or more copies of a single stranded or double stranded  
5 nucleic acid, or a portion thereof. In some embodiments, the methods provided herein can include a step of producing an amplified nucleic acid under isothermal or thermal variable conditions.

As used herein the terms “selectively” and “targeted” and the like, when used in reference to “amplifying” “amplification” (or grammatical equivalents), refers to  
10 preferentially amplifying a first nucleic acid in a sample compared to one or more other nucleic acids in the sample. The term can refer to producing one or more copies of the first nucleic acid and substantially no copies of the other nucleic acids. The term can also refer to producing a detectable amount of copies of the first nucleic acid and an undetectable (or insignificant) amount of copies of the other nucleic acids under a  
15 particular detection condition used.

Any suitable amplification methodology can be utilized to selectively or non-selectively amplify one or more nucleic acid molecules from an individual according to the methods and systems presented herein. It will be appreciated that any of the amplification methodologies described herein or generally known in the art can be  
20 utilized with target-specific primers to selectively amplify a nucleic acid molecule of interest. Suitable methods for selective amplification include, but are not limited to, the polymerase chain reaction (PCR), strand displacement amplification (SDA), transcription mediated amplification (TMA) and nucleic acid sequence based amplification (NASBA), as described in U.S. Patent No. 8,003,354, which is  
25 incorporated herein by reference in its entirety. The above amplification methods can be employed to selectively amplify one or more nucleic acids of interest. For example, PCR, including multiplex PCR, SDA, TMA, NASBA and the like can be utilized to selectively amplify one or more nucleic acids of interest. In such embodiments, primers directed specifically to the nucleic acid of interest are included in the  
30 amplification reaction.

Other suitable methods for amplification of nucleic acids can include oligonucleotide extension and ligation, rolling circle amplification (RCA) (Lizardi et al., Nat. Genet. 19:225-232 (1998), which is incorporated herein by reference) and oligonucleotide ligation assay (OLA) (See generally U.S. Pat. Nos. 7,582,420, 5,185,243, 5,679,524 and 5,573,907; EP 0 320 308 B1; EP 0 336 731 B1; EP 0 439 182 B1; WO 90/01069; WO 89/12696; and WO 89/09835, all of which are incorporated by reference) technologies. It will be appreciated that these amplification methodologies can be designed to selectively amplify a target nucleic acid of interest. For example, in some embodiments, the selective amplification method can include ligation probe  
5 amplification or oligonucleotide ligation assay (OLA) reactions that contain primers directed specifically to the nucleic acid of interest. In some embodiments, the selective amplification method can include a primer extension-ligation reaction that contains primers directed specifically to the nucleic acid of interest. As a non-limiting example of primer extension and ligation primers that can be specifically designed to amplify a  
10 nucleic acid of interest, the amplification can include primers used for the GoldenGate assay (Illumina, Inc., San Diego, CA).

Exemplary isothermal amplification methods that can be used in a method of the present disclosure include, but are not limited to, Multiple Displacement Amplification (MDA) as exemplified by, for example Dean et al., Proc. Natl. Acad. Sci. USA  
20 99:5261-66 (2002) or isothermal strand displacement nucleic acid amplification exemplified by, for example U.S. Pat. No. 6,214,587, each of which is incorporated herein by reference in its entirety. Other non-PCR-based methods that can be used in the present disclosure include, for example, strand displacement amplification (SDA) which is described in, for example Walker et al., Molecular Methods for Virus  
25 Detection, Academic Press, Inc., 1995; U.S. Pat. Nos. 5,455,166, and 5,130,238, and Walker et al., Nucl. Acids Res. 20:1691-96 (1992) or hyperbranched strand displacement amplification which is described in, for example Lage et al., Genome Research 13:294-307 (2003), each of which is incorporated herein by reference in its entirety. Isothermal amplification methods can be used with the strand-displacing Phi  
30 29 polymerase or Bst DNA polymerase large fragment, 5'->3'  $\text{exo}^-$  for random primer

amplification of genomic DNA. The use of these polymerases takes advantage of their high processivity and strand displacing activity. High processivity allows the polymerases to produce fragments that are 10-20 kb in length. As set forth above, smaller fragments can be produced under isothermal conditions using polymerases  
5 having low processivity and strand-displacing activity such as Klenow polymerase. Additional description of amplification reactions, conditions and components are set forth in detail in the disclosure of U.S. Patent No. 7,670,810, which is incorporated herein by reference in its entirety.

Another nucleic acid amplification method that is useful in the present  
10 disclosure is Tagged PCR which uses a population of two-domain primers having a constant 5' region followed by a random 3' region as described, for example, in Grothues et al. *Nucleic Acids Res.* 21(5):1321-2 (1993), incorporated herein by reference in its entirety. The first rounds of amplification are carried out to allow a multitude of initiations on heat denatured DNA based on individual hybridization from  
15 the randomly-synthesized 3' region. Due to the nature of the 3' region, the sites of initiation are contemplated to be random throughout the genome. Thereafter, the unbound primers can be removed and further replication can take place using primers complementary to the constant 5' region.

A further approach that can be used to amplify gDNA in connection with the  
20 methods of the present disclosure is degenerate oligonucleotide primed polymerase chain reaction (DOP-PCR) under conditions such as, but not limited to, those described by Cheung et al., *Proc. Natl. Acad. Sci. USA*, 93:14676-79 (1996) or U.S. Pat. No. 5,043,272, the disclosures of which are incorporated herein by reference in their entireties. Low amounts of gDNA, for example, 15 pg of human gDNA, can be  
25 amplified to levels that are conveniently detected in the methods of the present disclosure. Reaction conditions used in the methods of Cheung et al. can be selected for production of an amplified representative population of genome fragments having near complete coverage of the human genome. Furthermore modified versions of DOP-PCR, such as those described by Kittler et al. in a protocol known as LL-DOP-PCR (Long  
30 products from Low DNA quantities-DOP-PCR) can be used to amplify gDNA in

accordance with the present disclosure (Kittler et al., Anal. Biochem. 300:237-44 (2002), the disclosure of which is incorporated herein by reference in its entirety).

Primer-extension preamplification polymerase chain reaction (PEP-PCR) can also be used in a method of the present disclosure in order to amplify gDNA. Useful  
5 conditions for amplification of gDNA using PEP-PCR include, for example, those described in Casas et al., Biotechniques 20:219-25 (1996), incorporated herein by reference in its entirety.

In some embodiments, selective amplification can include a method to pull-down a nucleic acid of interest from a mixture of different nucleic acids. The pull-down  
10 can occur prior to or after amplification occurs. Pull-down methods are well known in the art, and can include, for example, nucleic acid pull-down using biotinylated probes or arrays of probes.

The present methods are not limited to any particular amplification technique and amplification techniques described herein are exemplary only with regards to  
15 methods and embodiments of the present disclosure.

The methods of the invention provide for a rapid test using “raw” sample input (e.g., unprocessed dried blood spot) and/or processed sample input (e.g., purified genomic DNA) that may be used to quickly and unequivocally confirm sample identity of WGS data.  
20

### **Sequencing Methods**

The methods described herein can be used in conjunction with a variety of nucleic acid sequencing techniques. Particularly applicable techniques are those wherein nucleic acids are attached at fixed locations in an array such that their relative  
25 positions do not change and wherein the array is repeatedly imaged. Embodiments in which images are obtained in different color channels, for example, coinciding with different labels used to distinguish one nucleotide base type from another are particularly applicable. In some embodiments, the process to determine the nucleotide sequence of a target nucleic acid can be an automated process. Preferred embodiments  
30 include sequencing-by-synthesis (“SBS”) techniques.

“Sequencing-by-synthesis (“SBS”) techniques” generally involve the enzymatic extension of a nascent nucleic acid strand through the iterative addition of nucleotides against a template strand. In traditional methods of SBS, a single nucleotide monomer may be provided to a target nucleotide in the presence of a polymerase in each delivery.

- 5 However, in the methods described herein, more than one type of nucleotide monomer can be provided to a target nucleic acid in the presence of a polymerase in a delivery.

SBS can utilize nucleotide monomers that have a terminator moiety or those that lack any terminator moieties. Methods utilizing nucleotide monomers lacking terminators include, for example, pyrosequencing and sequencing using  $\gamma$ -phosphate-labeled nucleotides, as set forth in further detail below. In methods using nucleotide monomers lacking terminators, the number of nucleotides added in each cycle is generally variable and dependent upon the template sequence and the mode of nucleotide delivery. For SBS techniques that utilize nucleotide monomers having a terminator moiety, the terminator can be effectively irreversible under the sequencing conditions used as is the case for traditional Sanger sequencing which utilizes dideoxynucleotides, or the terminator can be reversible as is the case for sequencing methods developed by Solexa (now Illumina, Inc.).

10  
15

SBS techniques can utilize nucleotide monomers that have a label moiety or those that lack a label moiety. Accordingly, incorporation events can be detected based on a characteristic of the label, such as fluorescence of the label; a characteristic of the nucleotide monomer such as molecular weight or charge; a byproduct of incorporation of the nucleotide, such as release of pyrophosphate; or the like. In embodiments, where two or more different nucleotides are present in a sequencing reagent, the different nucleotides can be distinguishable from each other, or alternatively, the two or more different labels can be the indistinguishable under the detection techniques being used. For example, the different nucleotides present in a sequencing reagent can have different labels and they can be distinguished using appropriate optics as exemplified by the sequencing methods developed by Solexa (now Illumina, Inc.).

20  
25

Preferred embodiments include pyrosequencing techniques. Pyrosequencing detects the release of inorganic pyrophosphate (PPi) as particular nucleotides are

30

incorporated into the nascent strand (Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) "Real-time DNA sequencing using detection of pyrophosphate release." *Analytical Biochemistry* 242(1), 84-9; Ronaghi, M. (2001) "Pyrosequencing sheds light on DNA sequencing." *Genome Res.* 11(1), 3-11; Ronaghi, M., Uhlen, M. and Nyren, P. (1998) "A sequencing method based on real-time pyrophosphate." *Science* 281(5375), 363; U.S. Pat. No. 6,210,891; U.S. Pat. No. 6,258,568 and U.S. Pat. No. 6,274,320, the disclosures of which are incorporated herein by reference in their entireties). In pyrosequencing, released PPi can be detected by being immediately converted to adenosine triphosphate (ATP) by ATP sulfurylase, and the level of ATP generated is detected via luciferase-produced photons. The nucleic acids to be sequenced can be attached to features in an array and the array can be imaged to capture the chemiluminescent signals that are produced due to incorporation of a nucleotides at the features of the array. An image can be obtained after the array is treated with a particular nucleotide type (e.g., A, T, C or G). Images obtained after addition of each nucleotide type will differ with regard to which features in the array are detected. These differences in the image reflect the different sequence content of the features on the array. However, the relative locations of each feature will remain unchanged in the images. The images can be stored, processed and analyzed using the methods set forth herein. For example, images obtained after treatment of the array with each different nucleotide type can be handled in the same way as exemplified herein for images obtained from different detection channels for reversible terminator-based sequencing methods.

In another exemplary type of SBS, cycle sequencing is accomplished by stepwise addition of reversible terminator nucleotides containing, for example, a cleavable or photobleachable dye label as described, for example, in International Patent Pub. No. WO 04/018497 and U.S. Patent 7,057,026, the disclosures of which are incorporated herein by reference. This approach is being commercialized by Solexa (now Illumina Inc.), and is also described in International Patent Pub. No. WO 91/06678 and International Patent Pub. No. WO 07/123,744, each of which is incorporated herein by reference. The availability of fluorescently-labeled terminators

in which both the termination can be reversed and the fluorescent label cleaved facilitates efficient cyclic reversible termination (CRT) sequencing. Polymerases can also be co-engineered to efficiently incorporate and extend from these modified nucleotides.

5            Preferably in reversible terminator-based sequencing embodiments, the labels do not substantially inhibit extension under SBS reaction conditions. However, the detection labels can be removable, for example, by cleavage or degradation. Images can be captured following incorporation of labels into arrayed nucleic acid features. In particular embodiments, each cycle involves simultaneous delivery of four different  
10    nucleotide types to the array and each nucleotide type has a spectrally distinct label. Four images can then be obtained, each using a detection channel that is selective for one of the four different labels. Alternatively, different nucleotide types can be added sequentially and an image of the array can be obtained between each addition step. In such embodiments each image will show nucleic acid features that have incorporated  
15    nucleotides of a particular type. Different features will be present or absent in the different images due the different sequence content of each feature. However, the relative position of the features will remain unchanged in the images. Images obtained from such reversible terminator-SBS methods can be stored, processed and analyzed as set forth herein. Following the image capture step, labels can be removed and  
20    reversible terminator moieties can be removed for subsequent cycles of nucleotide addition and detection. Removal of the labels after they have been detected in a particular cycle and prior to a subsequent cycle can provide the advantage of reducing background signal and crosstalk between cycles. Examples of useful labels and removal methods are set forth below.

25            In particular embodiments some or all of the nucleotide monomers can include reversible terminators. In such embodiments, reversible terminators/cleavable fluors can include fluor linked to the ribose moiety via a 3' ester linkage (Metzker, Genome Res. 15:1767-1776 (2005), which is incorporated herein by reference). Other approaches have separated the terminator chemistry from the cleavage of the  
30    fluorescence label (Ruparel et al., Proc Natl Acad Sci USA 102: 5932-7 (2005), which

is incorporated herein by reference in its entirety). Ruparel et al described the development of reversible terminators that used a small 3' allyl group to block extension, but could easily be deblocked by a short treatment with a palladium catalyst. The fluorophore was attached to the base via a photocleavable linker that could easily  
5 be cleaved by a 30 second exposure to long wavelength UV light. Thus, either disulfide reduction or photocleavage can be used as a cleavable linker. Another approach to reversible termination is the use of natural termination that ensues after placement of a bulky dye on a dNTP. The presence of a charged bulky dye on the dNTP can act as an effective terminator through steric and/or electrostatic hindrance. The presence of one  
10 incorporation event prevents further incorporations unless the dye is removed. Cleavage of the dye removes the fluor and effectively reverses the termination. Examples of modified nucleotides are also described in U.S. Patent 7,427,673, and U.S. Patent 7,057,026, the disclosures of which are incorporated herein by reference in their entireties.

15 Additional exemplary SBS systems and methods which can be utilized with the methods and systems described herein are described in U.S. Patent Pub. No. 2007/0166705, U.S. Patent Pub. No. 2006/0188901, U.S. Patent 7,057,026, U.S. Patent Pub. No. 2006/0240439, U.S. Patent Pub. No. 2006/0281109, International Patent Pub. No. WO 05/065814, U.S. Patent Pub. No. 2005/0100900, International Patent  
20 Pub. No. WO 06/064199, International Patent Pub. No. WO 07/010,251, U.S. Patent Pub. No. 2012/0270305 and U.S. Patent Pub. No. 2013/0260372, the disclosures of which are incorporated herein by reference in their entireties.

Some embodiments can utilize detection of four different nucleotides using fewer than four different labels. For example, SBS can be performed utilizing methods  
25 and systems described in the incorporated materials of U.S. Patent Pub. No. 2013/0079232. As a first example, a pair of nucleotide types can be detected at the same wavelength, but distinguished based on a difference in intensity for one member of the pair compared to the other, or based on a change to one member of the pair (e.g., via chemical modification, photochemical modification or physical modification) that  
30 causes apparent signal to appear or disappear compared to the signal detected for the

other member of the pair. As a second example, three of four different nucleotide types can be detected under particular conditions while a fourth nucleotide type lacks a label that is detectable under those conditions, or is minimally detected under those conditions (e.g., minimal detection due to background fluorescence, etc.). Incorporation of the first three nucleotide types into a nucleic acid can be determined based on presence of their respective signals and incorporation of the fourth nucleotide type into the nucleic acid can be determined based on absence or minimal detection of any signal. As a third example, one nucleotide type can include label(s) that are detected in two different channels, whereas other nucleotide types are detected in no more than one of the channels. The aforementioned three exemplary configurations are not considered mutually exclusive and can be used in various combinations. An exemplary embodiment that combines all three examples, is a fluorescent-based SBS method that uses a first nucleotide type that is detected in a first channel (e.g., dATP having a label that is detected in the first channel when excited by a first excitation wavelength), a second nucleotide type that is detected in a second channel (e.g., dCTP having a label that is detected in the second channel when excited by a second excitation wavelength), a third nucleotide type that is detected in both the first and the second channel (e.g., dTTP having at least one label that is detected in both channels when excited by the first and/or second excitation wavelength) and a fourth nucleotide type that lacks a label that is not, or minimally, detected in either channel (e.g., dGTP having no label).

Further, as described in the incorporated materials of U.S. Patent Pub. No. 2013/0079232, sequencing data can be obtained using a single channel. In such so-called one-dye sequencing approaches, the first nucleotide type is labeled but the label is removed after the first image is generated, and the second nucleotide type is labeled only after a first image is generated. The third nucleotide type retains its label in both the first and second images, and the fourth nucleotide type remains unlabeled in both images.

Some embodiments can utilize sequencing by ligation techniques. Such techniques utilize DNA ligase to incorporate oligonucleotides and identify the incorporation of such oligonucleotides. The oligonucleotides typically have different

labels that are correlated with the identity of a particular nucleotide in a sequence to which the oligonucleotides hybridize. As with other SBS methods, images can be obtained following treatment of an array of nucleic acid features with the labeled sequencing reagents. Each image will show nucleic acid features that have incorporated  
5 labels of a particular type. Different features will be present or absent in the different images due the different sequence content of each feature, but the relative position of the features will remain unchanged in the images. Images obtained from ligation-based sequencing methods can be stored, processed and analyzed as set forth herein. Exemplary SBS systems and methods which can be utilized with the methods and  
10 systems described herein are described in U.S. Patent 6,969,488, U.S. Patent 6,172,218, and U.S. Patent 6,306,597, the disclosures of which are incorporated herein by reference in their entirety.

Some embodiments can utilize nanopore sequencing (Deamer, D. W. & Akeson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing." Trends  
15 Biotechnol. 18, 147-151 (2000); Deamer, D. and D. Branton, "Characterization of nucleic acids by nanopore analysis". Acc. Chem. Res. 35:817-825 (2002); Li, J., M. Gershow, D. Stein, E. Brandin, and J. A. Golovchenko, "DNA molecules and configurations in a solid-state nanopore microscope" Nat. Mater. 2:611-615 (2003), the disclosures of which are incorporated herein by reference in their entirety). In such  
20 embodiments, the target nucleic acid passes through a nanopore. The nanopore can be a synthetic pore or biological membrane protein, such as  $\alpha$ -hemolysin. As the target nucleic acid passes through the nanopore, each base-pair can be identified by measuring fluctuations in the electrical conductance of the pore. (U.S. Patent 7,001,792; Soni, G. V. & Meller, "A. Progress toward ultrafast DNA sequencing using solid-state  
25 nanopores." Clin. Chem. 53, 1996-2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis." Nanomed. 2, 459-481 (2007); Cockroft, S. L., Chu, J., Amarin, M. & Ghadiri, M. R. "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution." J. Am. Chem. Soc. 130, 818-820 (2008), the disclosures of which are incorporated herein by reference in their entirety).  
30 Data obtained from nanopore sequencing can be stored, processed and analyzed as set

forth herein. In particular, the data can be treated as an image in accordance with the exemplary treatment of optical images and other images that is set forth herein.

Some embodiments can utilize methods involving the real-time monitoring of DNA polymerase activity. Nucleotide incorporations can be detected through  
5 fluorescence resonance energy transfer (FRET) interactions between a fluorophore-bearing polymerase and  $\gamma$ -phosphate-labeled nucleotides as described, for example, in U.S. Patent 7,329,492 and U.S. Patent 7,211,414 (each of which is incorporated herein by reference) or nucleotide incorporations can be detected with zero-mode waveguides as described, for example, in U.S. Patent 7,315,019 (which is incorporated herein by  
10 reference) and using fluorescent nucleotide analogs and engineered polymerases as described, for example, in U.S. Patent 7,405,281 and U.S. Patent Pub. No. 2008/0108082 (each of which is incorporated herein by reference). The illumination can be restricted to a zeptoliter-scale volume around a surface-tethered polymerase such that incorporation of fluorescently labeled nucleotides can be observed with low  
15 background (Levene, M. J. et al. "Zero-mode waveguides for single-molecule analysis at high concentrations." *Science* 299, 682-686 (2003); Lundquist, P. M. et al. "Parallel confocal detection of single molecules in real time." *Opt. Lett.* 33, 1026-1028 (2008); Korlach, J. et al. "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nano structures." *Proc. Natl.*  
20 *Acad. Sci. USA* 105, 1176-1181 (2008), the disclosures of which are incorporated herein by reference in their entireties). Images obtained from such methods can be stored, processed and analyzed as set forth herein.

Some SBS embodiments include detection of a proton released upon incorporation of a nucleotide into an extension product. For example, sequencing based  
25 on detection of released protons can use an electrical detector and associated techniques that are commercially available from Ion Torrent (Guilford, CT, a Life Technologies subsidiary) or sequencing methods and systems described in U.S. Patent Pub. No. 2009/0026082; U.S. Patent Pub. No. 2009/0127589; U.S. Patent Pub. No. 2010/0137143; or U.S. Patent Pub. No. 2010/0282617, each of which is incorporated  
30 herein by reference. Methods set forth herein for amplifying target nucleic acids using

kinetic exclusion can be readily applied to substrates used for detecting protons. More specifically, methods set forth herein can be used to produce clonal populations of amplicons that are used to detect protons.

The above SBS methods can be advantageously carried out in multiplex formats such that multiple different target nucleic acids are manipulated simultaneously. In particular embodiments, different target nucleic acids can be treated in a common reaction vessel or on a surface of a particular substrate. This allows convenient delivery of sequencing reagents, removal of unreacted reagents and detection of incorporation events in a multiplex manner. In embodiments using surface-bound target nucleic acids, the target nucleic acids can be in an array format. In an array format, the target nucleic acids can be typically bound to a surface in a spatially distinguishable manner. The target nucleic acids can be bound by direct covalent attachment, attachment to a bead or other particle or binding to a polymerase or other molecule that is attached to the surface. The array can include a single copy of a target nucleic acid at each site (also referred to as a feature) or multiple copies having the same sequence can be present at each site or feature. Multiple copies can be produced by amplification methods such as, bridge amplification or emulsion PCR as described in further detail below.

The methods set forth herein can use arrays having features at any of a variety of densities including, for example, at least about 10 features/cm<sup>2</sup>, 100 features/cm<sup>2</sup>, 500 features/cm<sup>2</sup>, 1,000 features/cm<sup>2</sup>, 5,000 features/cm<sup>2</sup>, 10,000 features/cm<sup>2</sup>, 50,000 features/cm<sup>2</sup>, 100,000 features/cm<sup>2</sup>, 1,000,000 features/cm<sup>2</sup>, 5,000,000 features/cm<sup>2</sup>, or higher.

An advantage of the methods set forth herein is that they provide for rapid and efficient detection of a plurality of target nucleic acid in parallel. Accordingly the present disclosure provides integrated systems capable of preparing and detecting nucleic acids using techniques known in the art such as those exemplified above. Thus, an integrated system of the present disclosure can include fluidic components capable of delivering amplification reagents and/or sequencing reagents to one or more immobilized DNA fragments, the system comprising components such as pumps,

valves, reservoirs, fluidic lines and the like. A flow cell can be configured and/or used in an integrated system for detection of target nucleic acids. Exemplary flow cells are described, for example, in U.S. Patent Pub. No. 2010/0111768 A1 and U.S. Patent App. No. 13/273,666, each of which is incorporated herein by reference. As exemplified for

5 flow cells, one or more of the fluidic components of an integrated system can be used for an amplification method and for a detection method. Taking a nucleic acid sequencing embodiment as an example, one or more of the fluidic components of an integrated system can be used for an amplification method set forth herein and for the delivery of sequencing reagents in a sequencing method such as those exemplified

10 above. Alternatively, an integrated system can include separate fluidic systems to carry out amplification methods and to carry out detection methods. Examples of integrated sequencing systems that are capable of creating amplified nucleic acids and also determining the sequence of the nucleic acids include, without limitation, the MiSeq™ platform (Illumina, Inc., San Diego, CA) and devices described in U.S. Patent App. No.

15 13/273,666, which is incorporated herein by reference.

### **TSCA Library Preparation and Sequencing for Sample Validation**

In one embodiment, the methods of the invention use the TruSeq® Custom Amplicon (referred to herein as “TSCA”) for preparation of targeted amplicon libraries

20 for subsequent sequencing and sample identification.

**Figure 1** illustrates a flow diagram of an example of a method 100 of using TruSeq® Custom Amplicon for preparation of targeted amplicon libraries and subsequent sequencing for sample identification. In this example, TruSeq® Custom Amplicon v1 is used. Method 100 includes, but is not limited to, the following steps.

25 At a step 110, an input sample is prepared. In one example, the input sample is a 3 mm punch from a DBS sample. In one embodiment, the DBS punch is placed directly in a well of a 96-well plate or other suitable reaction vessel such as a microcentrifuge tube or the like. In some embodiments, the punch is undergoes further manipulations, such as rinsing, soaking, incubating, or shaking, to process the sample

30 material and prepare it for downstream library preparation steps. For example, the

sample may be placed in a buffer for a suitable amount of time to allow further permeabilization of the sample material and free up the nucleic acid for better access to hybridization and amplification reagents. In some embodiments the sample is a punch from a blood spot. In some embodiments, the sample comprises FFPE.

5           At a step 115, a pool of upstream and downstream oligonucleotides specific to targeted SNPs is hybridized to genomic DNA on the DBS sample punch. In one example, the panel of oligonucleotide includes upstream and downstream oligonucleotides targeting 24 identity-informative SNPs. For example, an aliquot of the oligonucleotide pool and an aliquot of hybridization buffer are added directly to each  
10 well containing a DBS sample punch. The 96-well plate is placed on a heating block heated to about 96 °C, incubated for about 1 minute, and then the temperature is slowly cooled to about 40 °C (e.g., slowly cooled over about 2 hours). At the end of the hybridization incubation, the supernatant is removed from each well (leaving the DBS punch in the well) and transferred to a new 96-well plate.

15           At a step 120, the standard steps in the TSCA protocol for library preparation are performed (e.g., removal of unbound oligonucleotides, extension/ligation of bound oligonucleotides, PCR amplification of extension-ligation products, etc.). In one example, removal of unbound oligonucleotide is performed using a filter plate (i.e., TSCA v1).

20           At a step 125, an aliquot of the amplified DBS library samples are pooled and sequenced. Any one of a number of known sequencing methodologies as described hereinabove can be used to sequence the library. In some embodiments, the pooled library samples are sequenced using massively parallel sequencing, for example, on a MiSeq version 3 instrument (Illumina, Inc.)

25           At a step 130, the sequencing data is analyzed. For example, the sequencing data is analyzed using MiSeq Reporter tool using Smith-Waterman aligner and GATK variant calling at the 24 SNP positions.

          In another example (not shown), a TruSeq® Custom Amplicon protocol that includes a bead-based process for removal of unbound oligonucleotides (i.e., TSCA v2)  
30 is used to prepare targeted amplicon libraries for sequencing and sample identification.

To evaluate method 100 of Figure 1, blood samples from 10 healthy donors were used. The blood samples were purchased from Clinical Trials Laboratory Service. Upon receipt of each sample, dried blood spot samples were prepared by applying an aliquot (e.g., about 50  $\mu$ l to about 70  $\mu$ L) of each blood sample on individual filter paper cards (i.e., Guthrie cards) and drying the sample on the card. Table 2 below shows a list of the donor samples that were used to evaluate method 100 of Figure 1. Each sample is designated by a CS number, e.g., CS212, CS219, CS220, etc. *Italicized* CS numbers, i.e., CS219, CS224, and CS308 are blood samples from the same individual (i.e., Patient ID KD241283) provided at different time points and used as a reference individual. All blood spots were less than 6 months old ("Date of blood arrival"). All blood samples were within a "normal" range (e.g., 3,000-10,000/ $\mu$ L in healthy individuals) for white blood cell counts ("WBC count") although the WBC counts differ between samples. For some donor blood samples, genomic DNA (indicated by "x" in "DNA stored" column) was also isolated from the blood sample using a QiaAmp kit and extraction protocol (available from Qiagen). Genomic DNA was used as a comparative control for library preparation and SNP calling.

| <b>Table 2. Donor blood samples</b> |                   |                              |                                                   |                                   |                       |
|-------------------------------------|-------------------|------------------------------|---------------------------------------------------|-----------------------------------|-----------------------|
| <b>Blood Spots</b>                  |                   |                              |                                                   |                                   |                       |
| <b>CS Number</b>                    | <b>Patient ID</b> | <b>Date of blood arrival</b> | <b>WBC count<br/>(<math>\times 10^9/L</math>)</b> | <b>Blood stored at<br/>-80 °C</b> | <b>gDNA extracted</b> |
| CS212                               | DP090486          | 16/12/2014                   | 13.9                                              | -                                 | -                     |
| <i>CS219</i>                        | <i>KD241283</i>   | <i>14/01/2015</i>            | <i>9.14</i>                                       | <i>x</i>                          | <i>x</i>              |
| CS220                               | B1070293          | 20/01/2015                   | 7.22                                              | x                                 | x                     |
| CS221                               | VP180886          | 20/01/2015                   | 6.06                                              | x                                 | x                     |
| CS222                               | RR310794          | 20/01/2015                   | 6.57                                              | x                                 | x                     |
| CS223                               | KK150693          | 20/01/2015                   | 4.51                                              | x                                 | x                     |
| CS224                               | SK020371          | 20/01/2015                   | 5.32                                              | x                                 | x                     |
| CS225                               | NS061189          | 20/01/2015                   | 6.69                                              | x                                 | x                     |
| <i>CS244</i>                        | <i>KD241283</i>   | <i>18/03/2015</i>            | <i>9.3*</i>                                       | <i>x</i>                          | <i>x</i>              |
| CS307                               | DS220595          | 14/04/2015                   | -                                                 | x                                 | -                     |
| <i>CS308</i>                        | <i>KD241283</i>   | <i>28/04/2015</i>            | -                                                 | <i>x</i>                          | <i>x</i>              |
| CS313                               | KM191186          | 19/05/2015                   | 8.3*                                              | x                                 | -                     |

\*In house WBC count

Targeted amplicon libraries were prepared from a 3 mm punch of each DBS sample. A 3 mm DBS punch contains about 200 ng of DNA. As a comparison, targeted amplicon libraries were also prepared from 100 ng of each genomic DNA sample (i.e., samples CS219, CS220, CS221, CS222, CS223, CS224, and CS225).

5        **Figure 2** shows a Agilent TapeStation image 200 of a gel of the PCR amplification products generated from each DBS punch and genomic DNA samples using method 100 of Figure 1. The DBS samples and genomic DNA (gDNA, 100 ng) control samples are as described with reference to Table 2. A bracket indicates the position of the bands representing the PCR amplification products. The lane labeled  
10        “Promega” is a positive control using a control DNA from Promega. The lane labeled “Blank” is a negative control using a filter punch that does not include DNA. The data show amplification products were obtained directly from DBS samples. Amplification products were also obtained from genomic DNA.

**Figure 3** shows a screenshot 300 of illumina sequence analysis viewer (SAV)  
15        software summarizing quality metrics for the MiSeq sequence run using the pooled DBS library samples of Figure 2. In this sequencing run, 35 million reads passed filter (about 1.5 million reads per sample). This yielded at least about 5,000x coverage for each identity SNP in the oligonucleotide panel.

**Figure 4** shows a plot 400 of the percent reads (passing filter) identified as a  
20        function of index number for the sequencing run of Figure 3. The data show even representation of each sample in the pooled sequencing library. The circled points on plot 400 are the negative controls.

The sequencing data was analyzed on MiSeq Reporter using Smith-Waterman aligner and the GATK variant caller. Variant calls at the 24 SNPs were scored relative  
25        to the hg19 reference. The calls were homozygous reference allele (“ref”), heterozygous (“het”), or homozygous alternative allele (“alt”). Calls from DBS samples, genomic DNA (extracted from the blood sample), and whole genome sequence (WGS) data (where available) were compared. **Figure 5** shows a table 500 of the SNP calls for 7 donor samples (i.e., CS219, CS220, CS221, CS222, CS223, CS224,  
30        and CS225) for both genomic DNA (gDNA) and DBS sample (“spot”). The data show

that identical genotype calls were made from libraries prepared from genomic DNA extracted from a blood sample and libraries prepared from the corresponding DNA on the dried blood spot sample.

For each donor sample (i.e., genomic DNA and DBS samples) shown table 500 of Figure 5, the data for all 24 identity SNP positions was collapsed into a single metric and used to determine identity. For example, for any given pair of donor samples, identity by state (IBS) can be observed at a given SNP and the relatedness of the samples determined. A DNA segment (e.g., SNP allele) is identical by state in two or more individuals (donors) if they have identical nucleotide sequences in this segment.

**Figures 6A and 6B** show a color key and histogram 600 and a plot 605 of identity by state (IBS) calculations, respectively, for the donor samples shown in table 500 of Figure 5. Referring to Figure 6A, the red color indicates 100% identity (value = 1), which means that all 24 SNPs for any given pair of donor samples are the same. The color scale transitions to green which indicates 50% (value = 0.5) of the SNPs for any given pair of donor samples are the same. Referring to Figure 6B, plot 605 compares SNPs in genomic DNA (“Extracted gDNA”) and DBS (“Blood Spots”) samples based on identity by state. On the diagonal (red boxes), all 24 SNPs in the corresponding genomic DNA and DBS samples are the same, e.g., CS219\_gDNA compared to CS219\_spot and CS220\_gDNA compared to CS220\_spot, etc. In contrast, the 24 SNPs in unrelated samples, e.g., CS219\_gDNA compared to CS220\_spot, are not the same as indicated by the green box, which indicates that the samples are not the same.

**Figures 7A and 7B** show a color key and histogram 700 and a plot 705 of identity by state (IBS) calculations, respectively, and show an example of discriminating DBS samples of Table 2 based on SNP data. Color key and histogram 700 is as described with reference to Figure 6A. Referring now to Figure 7B, plot 705 compares SNPs between DBS (“spot”) samples based on identity by state. On the diagonal (red boxes), all 24 SNPs in the matching DBS samples are the same, e.g., CS212\_spot compared to CS212\_spot, CS219\_spot compared to CS219\_spot, etc. The dashed arrows at CS219\_spot, CS244-spot, and CS308\_spot indicate samples from the same donor that were taken at different time points. All SNPs in samples

CS219\_spot, CS244-spot, and CS308\_spot are the same (red boxes), e.g., CS219\_spot compared to CS308\_spot, CS219\_spot compared to CS244\_spot, etc., and indicated that the samples are from the same individual. The solid arrow at CS244\_WGS indicates whole genome sequencing data; this sample was used to check concordance with the whole genome vs TSCA amplicon data.

The data show that method 100 of Figure 1 may be used to distinguish samples from different individuals and to correctly identify samples from the same individual taken at different time points.

In another example, targeted DNA amplicons may be prepared directly from FFPE samples using a TSCA library preparation protocol. To demonstrate preparation of targeted DNA amplicons directly for FFPE samples, slices (e.g., a 10  $\mu$ m slice) from a breast tumor FFPE sample and a gastric tumor FFPE sample were used. In this example, the targeted DNA panel was a colorectal cancer panel (Illumina). Slices of each breast tumor and gastric tumor samples were placed directly in wells of a 96-well plate and an aliquot of hybridization buffer and an aliquot of an upstream and downstream oligonucleotide pool specific to targeted genes is added (e.g., total reaction volume of 50  $\mu$ L) to each well. After the hybridization period and ligation/ extension reaction, the targeted DNA was amplified using 30 cycles of PCR. Genomic DNA was also prepared from breast and gastric FFPE samples and used as controls in the TSCA amplification protocol (i.e., input genomic DNA = 100 ng).

**Figure 8** shows an Agilent TapeStation image 800 of a gel of the PCR amplification products from FFPE and genomic DNA samples in pool A and an Agilent TapeStation image 805 of the PCR amplification products from FFPE and genomic DNA samples in pool B, wherein pool A and B are separate oligo pools that target opposite strands of the respective DNA target, which provide two sets of equivalent data useful for excluding errors induced through PCR, sequencing or FFPE induced base modifications. A bracket in Figure 8 indicates the position of the bands representing the PCR amplification products. The lane labeled “No DNA” is a negative control. The data show amplification products were obtained directly from FFPE samples. Amplification products were also obtained from genomic DNA.

### **Multiplex Targeted PCR Amplification Libraries and Sequencing for Sample Validation**

In another embodiment, the methods of the invention use multiplex targeted  
5 PCR amplification for preparation of genomic amplicon libraries for subsequent  
sequencing and sample validation.

**Figure 9** illustrates a flow diagram of an example of method 900 of using  
multiplex targeted PCR amplification for preparation of targeted amplicon libraries and  
subsequent sequencing for sample identification. Method 900 includes, but is not  
10 limited to, the following steps.

At a step 910, an input sample is prepared. In one example, the input sample is  
a 3 mm punch from a DBS sample. The DBS punch is placed directly in the well of a  
96-well plate.

At a step 915, multiplex targeted PCR amplification is performed on the DBS  
15 sample. For example, a targeted primer mix, PCR reagents, and Phusion Hot Start II  
High-Fidelity DNA Polymerase (Life Technologies) are added to each well with a DBS  
punch therein. The targeted primer mix includes primer pairs for 45 identity SNPs that  
include index adapters (i.e., 45-plex ID SNP subset from ForenSeq set (Illumina, Inc)).

At a step 920, the amplified DNA is pooled and loaded onto a flow cell prepared  
20 with capture probes. The capture probes are specific for the amplified target sequences.  
The targeted DNA sequences are captured on the flow cell by hybridization to the  
capture probes on the flow cell surface.

At a step 925, captured target sequences are clonally amplified on the flow cell  
surface.

25 At a step 930, the amplified DNA is sequenced. For example, the amplified  
DNA is sequenced on a MiSeq V3 instrument using 2x 76 cycles of sequencing.

At a step 935, the sequencing data is analyzed. For example, the sequencing  
data is analyzed using the Burrow-Wheeler Aligner (BWA) in the PCR amplicon work  
flow in MiSeq Reporter.

Method 900 was evaluated using the blood samples described above with reference to Table 2. Targeted amplicon libraries were prepared from a 3 mm punch of each DBS sample. A 3 mm DBS punch contains about 200 ng of DNA. Targeted amplicon libraries were also prepared from 1 ng of genomic DNA from samples CS219, CS220, CS221, CS222, CS223, CS224, and CS225.

**Figures 10A and 10B** show a BioAnalyzer image 1000 of a gel of the PCR amplification products generated from gDNA and a BioAnalyzer image 1005 of the PCR amplification products generated from each DBS punch, respectively, using method 900 of Figure 9. The DBS samples and genomic DNA (gDNA, 1 ng) control samples are as described above with reference to Table 2. A bracket indicates the position of the bands representing the PCR amplification products. Referring to Figure 10A, the lane labeled “-ve” is a negative control that does not include template DNA. Lanes 12878, 12877, 12884, and 12885 are control samples from the Platinum Genomes family (i.e., father, mother, and 2 children). Some amplification products were observed in the negative control sample. The data show amplification products were obtained directly from DBS samples. Amplification products were also obtained from genomic DNA.

**Figure 11** shows a screenshot 1100 of illumina sequence analysis viewer (SAV) software summarizing quality metrics for the MiSeq sequence run using the DBS and genomic amplicon libraries of Figures 10A and 10B. In this sequencing run, one surface of the flow cell was scanned, 3.71 million reads passed filter for about 10s – 1000s x coverage (a low coverage run) for each identity SNP in the oligonucleotide panel.

**Figure 12** shows a plot 1200 of the percent reads (passing filter) identified as a function of index number for the sequencing run of Figure 11. Reads from the DBS samples (“spots”) are above the line and reads from the genomic DNA samples (“gDNA”) are below the line. The discrepancy in read representation for DBS and genomic DNA samples may be due to a lower efficiency in sample pooling and/or normalization.

The sequencing data was analyzed on MiSeq Reporter using the Burrow-Wheeler Aligner (BWA) in the PCR amplicon workflow in MiSeq Reporter.

For each donor sample (i.e., genomic DNA and DBS samples), the data for all 45 identity SNP positions was collapsed into a single metric and used to determine  
5 identity.

**Figures 13A and 13B** show a color key and histogram 1300 and a plot 1305 of identity by state (IBS) calculations, respectively, for the donor samples of Table 2 used to evaluate method 900 of Figure 9. Referring to Figure 13A, the red color indicates 100% identity (value = 1), which means that all 45 SNPs for any given pair of donor  
10 samples are the same. The color scale transitions to green which indicates 50% (value = 0.5) of the SNPs for any given pair of donor samples are the same. Referring to Figure 13B, plot 1305 compares SNPs in genomic DNA (“Extracted gDNA”) and DBS (“Blood Spots”) samples based on identity by state. On the diagonal (red boxes), all 45 SNPs in the corresponding genomic DNA and DBS samples are the same, e.g.,  
15 CS219\_dna compared to CS219\_spot and CS220\_dna compared to CS220\_spot, etc., which indicates that the sample identity is called correctly. In contrast, the 45 SNPs in unrelated samples, e.g., CS219\_dna compared to CS220\_spot, are not the same as indicated by the green box, which indicates that the samples are not the same. Note that for CS224\_dna compared to CS224\_spot, the box is not as red as the other matched  
20 genomic DNA/ DBS samples which may be due to lower sequencing coverage of some amplicons.

**Figures 14A and 14B** show a color key and histogram 1400 and a plot 1405 of identity by state (IBS) calculations, respectively, and show an example of discriminating DBS samples of Table 2 based on SNP data. Color key and histogram  
25 1400 is as described with reference to Figure 13A. Referring now to Figure 14B, plot 1405 compares SNPs between DBS (“spot”) and genomic DNA (“dna”) samples based on identity by state. On the diagonal (red boxes), all 45 SNPs in the matching DBS samples are the same, e.g., CS212spot compared to CS212spot, CS219dna compared to CS219dna, etc. The dashed arrows at CS219dna, CS219spot, CS244spot, and  
30 CS308spot indicate samples from the same donor (reference sample) that were taken at

different time points. All 45 SNPs in the reference samples CS219, CS244, and CS308 are identified as the same (red boxes), e.g., CS219spot compared to CS308spot, CS219dna compared to CS244spot, etc., and indicated that the samples are from the same individual. The solid arrow at CS244dna indicates whole genome sequencing data; this sample was used to check concordance with the whole genome vs PCR amplicon data.

The data show that method 900 of Figure 9 may be used to distinguish samples from different individuals and to correctly identify samples from the same individual taken at different time points.

10

### **High Resolution Sample Discrimination**

In another embodiment, the methods of the invention are used to discriminate between samples from closely related individuals. In one example, the closely related individuals are siblings. In another example, the closely related individuals are a family trio, wherein both parents and at least one child are afflicted, for example, with a disease or condition.

To evaluate the efficacy of the methods of the invention in high resolution sample discrimination, a set of samples from the Platinum Genomes family was used. The Platinum Genomes is a high confidence, “platinum” quality reference variant call set that was generated by sequencing a large family to high depth using a PCR-Free sample prep to maximize variant calling sensitivity as described by Eberle et al. (2016) bioRxiv doi: 10.1101/055541, the content of which is incorporated herein by reference in its entirety. Sample discrimination was performed using method 100 of Figure 1 and the 24 identity SNP panel.

25       **Figure 15** shows a family tree diagram 1500 of the Platinum Genomes family. The arrows indicate family members whose genomic DNA was used in the evaluation (i.e., father, mother, 4 female children, and 5 male children).

30       **Figures 16A and 16B** show a color key and histogram 1600 and a plot 1605 of identity by state (IBS) calculations, respectively, and show an example of high resolution genomic DNA sample discrimination based on SNP data. The genomic

DNA samples are as described with reference to Figure 15. Referring to Figure 16A, the red color indicates 100% identity (value = 1), which means that all 24 SNPs for any given pair of donor samples are the same. The color scale transitions to green which indicates 50% (value = 0.5) of the SNPs for any given pair of donor samples are the same. Referring now to Figure 16B, plot 1605 compares SNPs between genomic samples based on identity by state. On the diagonal (red boxes), all 24 SNPs in the matching genomic samples are the same, e.g., NA12877 compared to NA12877, NA12878 compared to NA12878, etc. The data show that method 100 of Figure 1 may be used to distinguish samples from closely related individuals.

10

Throughout this application various publications, patents and/or patent applications have been referenced. The disclosure of these publications in their entireties is hereby incorporated by reference in this application.

The term comprising is intended herein to be open-ended, including not only the recited elements, but further encompassing any additional elements. A number of embodiments are set forth in this description and in the incorporated materials of U.S. Provisional Patent Application No. 61/189,063 filed on July 6, 2015. Nevertheless, it will be understood that various modifications may be made. Accordingly, other embodiments are within the scope of the following claims.

20

What is claimed is:

1. A method for obtaining nucleic acid sequence information from a FFPE sample comprising:

(a) providing a FFPE sample comprising different target nucleic acids embedded within a preserved tissue;

(b) contacting the FFPE sample with a plurality of different probe sets to form hybridization complexes with the different target nucleic acids;

(c) amplifying the nucleic acid from the FFPE samples to produce amplicons;

wherein there is no purification of the nucleic acid from the FFPE sample prior to the contacting step (b); and

(d) obtaining nucleic acid sequence information for a plurality of the amplicons.

2. The method of claim 1, wherein there is no purification of the nucleic acid from the FFPE sample prior to the amplifying in step (c).

3. The method of claim 1, wherein the amplifying in step (c) comprises polymerase chain reaction using at least two amplification primers that are specific for a portion of the sample genome.

4. The method of claim 1, wherein amplifying in step (c) comprises extension and ligation of two probes to form amplification templates.

5. The method of claim 4, wherein the supernatant comprising solution-phase hybridization complexes is further contacted with a solid support to form immobilized hybridization complexes when contacted with the extension enzyme and the nucleotides.

6. The method of claim 5, wherein the solid support comprises beads.

7. The method of claim 5, wherein the solid support comprises a filter plate.

8. The method of claim 4, wherein each of the different target nucleic acids comprises, from 3' to 5': contiguous first, second, and third target domains and each probe set comprises: (i) a first probe comprising, from 5' to 3': a first priming sequence and a sequence that is substantially complementary to a first target domain; and (ii) a

second probe comprising 5' to 3': a sequence substantially complementary to a third target domain, and a second priming sequence.

9. The method of claim 1, comprising, prior to step (c), a step of collecting a supernatant comprising solution-phase hybridization complexes from the FFPE sample.

10. The method of claim 1, wherein obtaining nucleic acid sequence information comprises massively parallel sequencing.

11. The method of claim 1, wherein obtaining nucleic acid sequence information comprises detecting the amplicons on the surface of a nucleic acid array.

12. The method of claim 1, wherein the plurality of probe set comprises at least at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 different probe sets.

13. The method of claim 1, wherein the plurality of probe set comprises a plurality of probes configured to selectively hybridize to polymorphic regions that are informative of the identity of the source of the sample.

14. The method of claim 1, wherein the plurality of probe set comprises a plurality of probes configured to selectively hybridize to regions that comprise cancer-associated polymorphisms.

15. The method of claim 1, wherein the sample is a human sample.

16. The method of claim 1, wherein the sample comprises tumor tissue.

17. A method for amplification of nucleic acid from a FFPE sample comprising:

(a) providing a FFPE sample comprising nucleic acid embedded within a preserved tissue, the nucleic acid having, from 3' to 5': contiguous first, second, and third target domains;

(b) contacting the FFPE sample with a plurality of different probe sets to form hybridization complexes with the different target nucleic acids, wherein each probe set comprises: (i) a first probe comprising, from 5' to 3': a first priming sequence and a sequence that is substantially complementary to the first

target domain; and (ii) a second probe comprising 5' to 3': a sequence substantially complementary to the third target domain, and a second priming sequence;

(c) contacting the hybridization complexes with an extension enzyme and nucleotides, wherein the first probes are extended along the second target domains of hybridization complexes formed in (b);

(d) ligating the extended first probes to the second probes to form amplification templates; and

(e) amplifying the amplification templates with first and second primers that are complementary to the first priming sequence and the second priming sequence to produce amplicons and obtaining nucleic acid sequence information for a plurality of the amplicons.

18. The method of claim 17, comprising, prior to step (c), a step of collecting a supernatant comprising solution-phase hybridization complexes from the FFPE sample.

19. The method of claim 18, wherein the supernatant comprising solution-phase hybridization complexes is further contacted with a solid support to form immobilized hybridization complexes when contacted with the extension enzyme and the nucleotides.

20. The method of claim 19, wherein the solid support comprises beads.

21. The method of claim 19, wherein the solid support comprises a filter plate.

22. The method of claim 17, wherein obtaining nucleic acid sequence information comprises massively parallel sequencing.

23. The method of claim 17, wherein obtaining nucleic acid sequence information comprises detecting the amplicons on the surface of a nucleic acid array.

24. The method of claim 17, wherein the plurality of probe set comprises at least at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 different probe sets.

25. The method of claim 17, wherein the plurality of probe set comprises a plurality of probes configured to selectively hybridize to polymorphic regions that are informative of the identity of the source of the sample.

5 26. The method of claim 17, wherein the plurality of probe set comprises a plurality of probes configured to selectively hybridize to regions that comprise cancer-associated polymorphisms.

27. The method of claim 17, wherein the sample is a human sample.

28. The method of claim 17, wherein the sample comprises tumor tissue.

29. The method of claim 17, wherein the sample comprises normal tissue.

10 30. A method for obtaining nucleic acid sequence information from a biological sample comprising:

(a) providing a biological sample comprising different target nucleic acids;

15 (b) contacting the biological sample with a plurality of different probe sets to form hybridization complexes with the different target nucleic acids;

(c) amplifying the nucleic acid from the biological samples to produce amplicons;

wherein there is no purification of the nucleic acid from the biological sample prior to the contacting step (b); and

20 (d) obtaining nucleic acid sequence information for a plurality of portions of the amplified sample.

31. The method of claim 30, wherein there is no purification of the nucleic acid from the biological sample prior to the amplifying in step (c).

25 32. The method of claim 30, wherein the amplifying in step (c) comprises polymerase chain reaction using at least two amplification primers that are specific for a portion of the sample genome.

33. The method of claim 30, wherein amplifying in step (c) comprises extension and ligation of two probes to form amplification templates.

30 34. The method of claim 33, wherein the supernatant comprising solution-phase hybridization complexes is further contacted with a solid support to form immobilized hybridization complexes when contacted with the extension enzyme and the nucleotides.

35. The method of claim 34, wherein the solid support comprises beads.

36. The method of claim 34, wherein the solid support comprises a filter plate.

37. The method of claim 33, wherein each of the different target nucleic acids comprises, from 3' to 5': contiguous first, second, and third target domains and each probe set comprises: (i) a first probe comprising, from 5' to 3': a first priming sequence and a sequence that is substantially complementary to a first target domain; and (ii) a second probe comprising 5' to 3': a sequence substantially complementary to a third target domain, and a second priming sequence.

38. The method of claim 30, comprising, prior to step (c), a step of collecting a supernatant comprising solution-phase hybridization complexes from the biological sample.

39. The method of claim 30, wherein obtaining nucleic acid sequence information comprises massively parallel sequencing.

40. The method of claim 30, wherein obtaining nucleic acid sequence information comprises detecting the amplicons on the surface of a nucleic acid array.

41. The method of claim 30, wherein the plurality of probe set comprises at least at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 different probe sets.

42. The method of claim 30, wherein the plurality of probe set comprises a plurality of probes configured to selectively hybridize to polymorphic regions that are informative of the identity of the source of the sample.

43. The method of claim 30, wherein the plurality of probe set comprises a plurality of probes configured to selectively hybridize to regions that comprise cancer-associated polymorphisms.

44. The method of claim 30, wherein the sample is a human sample.

45. The method of claim 30, wherein the sample comprises tumor tissue.

46. The method of claim 30, wherein the sample comprises normal tissue.
47. The method of claim 30, wherein the sample is a blood sample.
48. The method of claim 47, wherein the sample comprises dried blood on a  
5 porous solid surface.
49. The method of claim 47, wherein the porous solid surface comprises filter paper.
50. A method for nucleic acid sample identification comprising:
- (a) providing a nucleic acid-containing cellular sample;
- 10 (b) lysing cells of the sample with a lysis reagent to liberate nucleic acid from within the cells of the cellular sample, thereby forming a lysate;
- (c) amplifying the nucleic acid from the lysed samples;
- wherein there is no purification of the nucleic acid from the lysate prior to beginning the amplification step (c); and
- 15 (d) obtaining nucleic acid sequence information for a plurality of portions of the amplified sample, and comparing the sequence information to a second set of sequence information.
51. The method of claim 1, wherein the nucleic acid is DNA.
52. The method of claim 1, wherein the sample is a blood sample.
- 20 53. The method of claim 1, wherein the sample comprises dried blood.
54. The method of claim 1, wherein the sample comprises a FFPE tissue sample.
55. The method of claim 1, wherein the second set of sequence information comprises a whole genome sequence.
- 25 56. The method of claim 1, wherein the second set of sequence information comprises exome sequence information.
57. The method of claim 1, wherein amplifying comprises a targeted amplification reaction.
58. The method of claim 57, wherein the targeted amplification reaction  
30 comprises extension and ligation of two probes.

59. The method of claim 8, wherein the targeted amplification reaction comprises polymerase chain reaction using at least two amplification primers that are specific for a portion of the sample genome.

60. A method of tracking the identity of a biological sample during different  
5 stages of sample processing, comprising:

(a) providing a nucleic acid-containing cellular sample;  
(b) separating a portion of the sample into a first portion and a second  
portion and obtaining a first set of nucleic acid sequence information from the first  
portion the biological sample according to any of claims 1-59, wherein the first set of  
10 nucleic acid sequence information comprises identity informative sequence information;

(c) purifying nucleic acid from the second portion and obtaining a second set  
of sequence information; and

(d) using computer-assisted logic, comparing the identity informative  
sequence information from the first set of nucleic acid sequence information sequence  
15 information to the second set of sequence information to confirm that the first and  
second sets of sequence information were obtained from the same source.

61. The method of claim 60, wherein the identity informative sequence  
information comprises SNP genotype information for at least 10, 11, 12, 13, 14, 15, 16,  
17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39,  
20 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least  
100 unique SNPs.

62. The method of claim 60, wherein the second set of sequence information  
comprises SNP genotype information for at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,  
20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42,  
25 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 unique  
SNPs.

63. The method of claim 60, wherein the nucleic acid is DNA.

64. The method of claim 60, wherein the sample is a blood sample.

65. The method of claim 60, wherein the sample comprises dried blood.

30 66. The method of claim 65, wherein the sample comprises dried blood on a  
porous solid surface.

67. The method of claim 60, wherein the sample comprises a FFPE tissue sample.

68. The method of claim 60, wherein the second set of sequence information comprises a whole genome sequence.

5 69. The method of claim 60, wherein the second set of sequence information comprises exome sequence information.

70. A method of confirming the source of two different biological samples comprising:

- (a) providing a first nucleic acid-containing cellular sample;
- 10 (b) obtaining a first set of nucleic acid sequence information from the first portion the biological sample according to any of claims 1-59, wherein the first set of nucleic acid sequence information comprises identity informative sequence information;
- (c) providing a second nucleic acid sample comprising purified nucleic acid and obtaining a second set of sequence information; and
- 15 (d) using computer-assisted logic, comparing the identity informative sequence information from the first set of nucleic acid sequence information sequence information to the second set of sequence information to confirm that the first and second sets of sequence information were obtained from the same individual.

71. The method of claim 70, wherein the identity informative sequence  
20 information comprises SNP genotype information for at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 unique SNPs.

72. The method of claim 70, wherein the second set of sequence information  
25 comprises SNP genotype information for at least 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 60, 70, 80, 90, or at least 100 unique SNPs.

73. The method of claim 70, wherein the nucleic acid is DNA.

30 74. The method of claim 70, wherein the sample is a blood sample.

75. The method of claim 70, wherein the sample comprises dried blood.

76. The method of claim 75, wherein the sample comprises dried blood on a porous solid surface.

77. The method of claim 70, wherein the sample comprises a FFPE tissue sample.

5        78. The method of claim 70, wherein the second set of sequence information comprises a whole genome sequence.

79. The method of claim 70, wherein the second set of sequence information comprises exome sequence information.

10

1/16

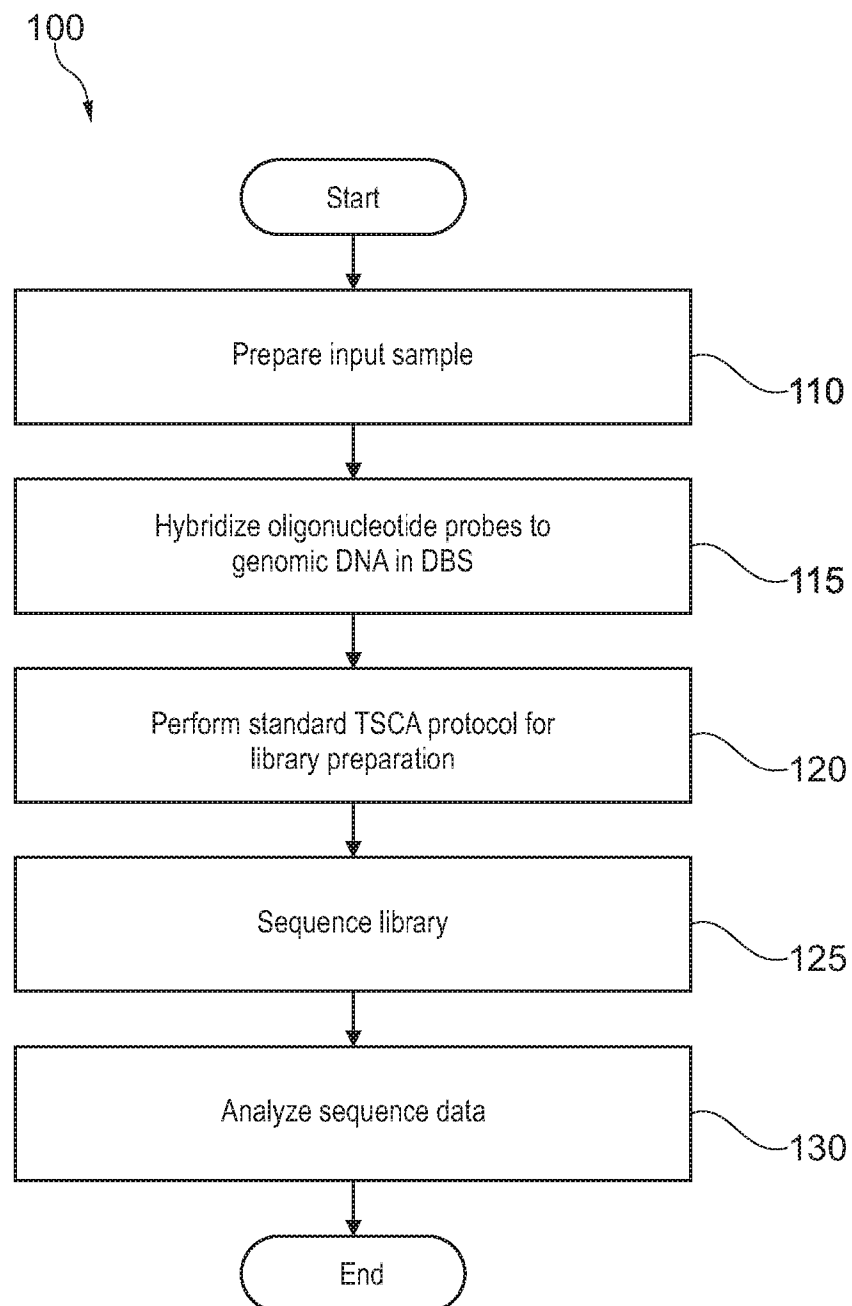


Fig. 1

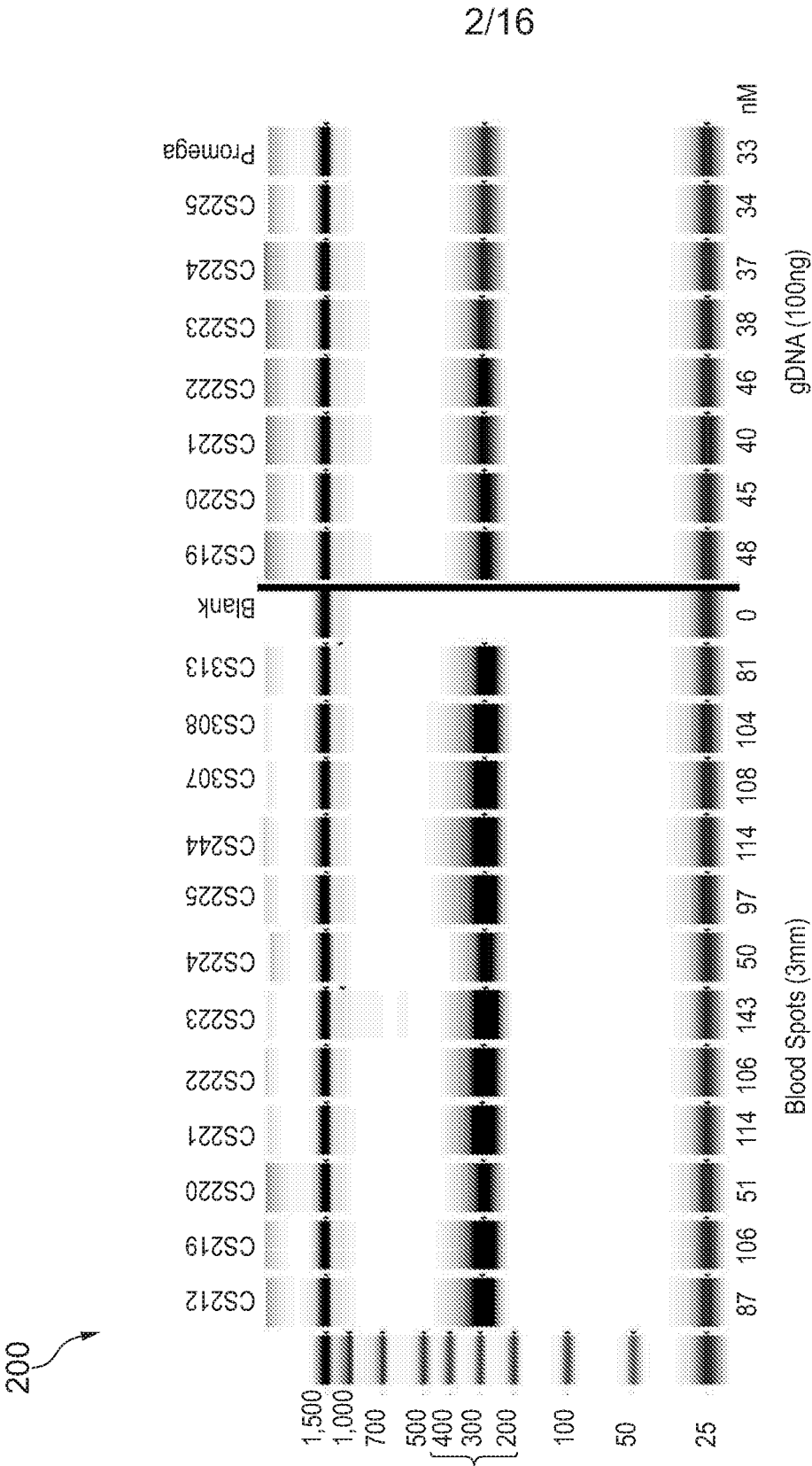


Fig. 2

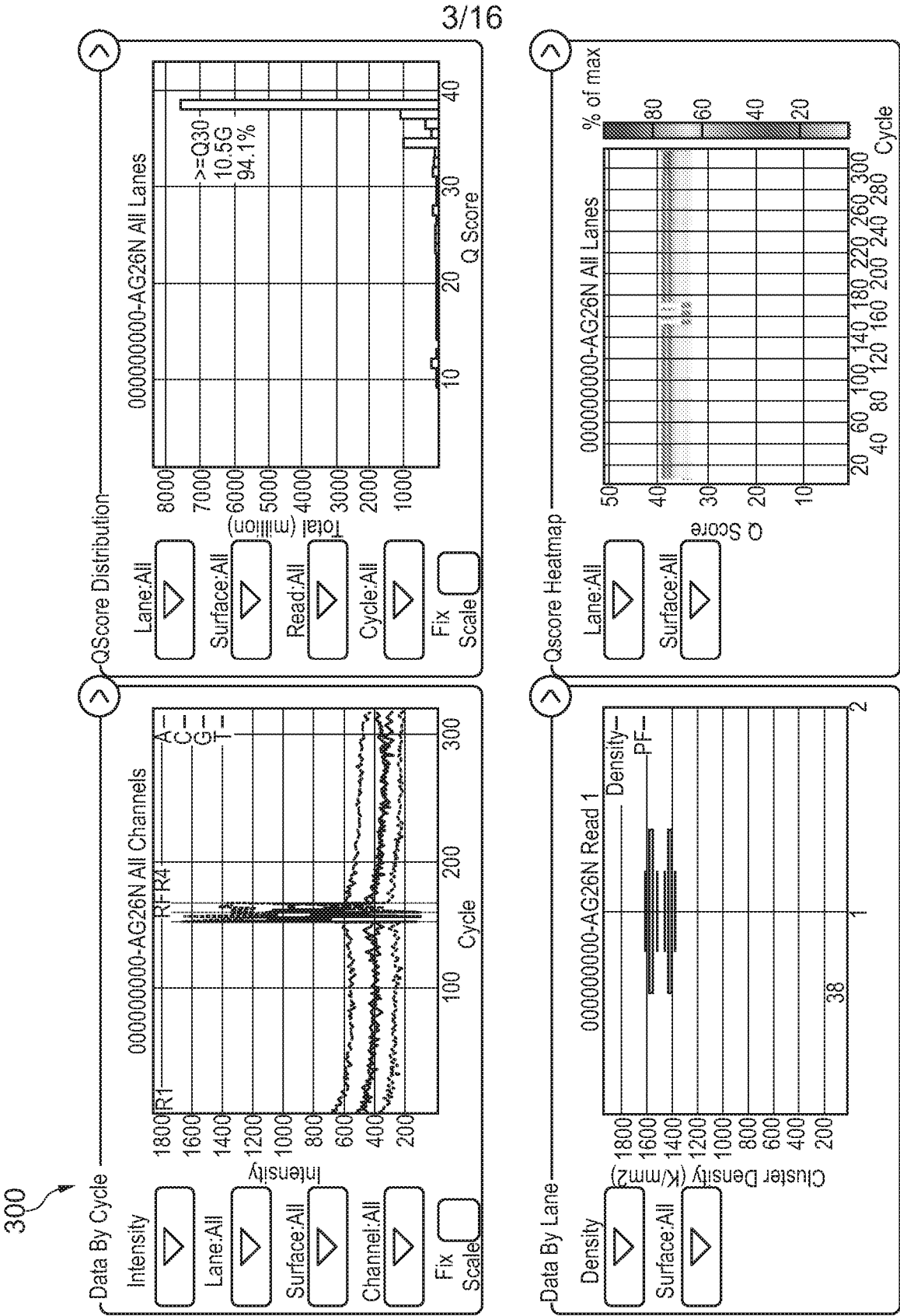


Fig. 3

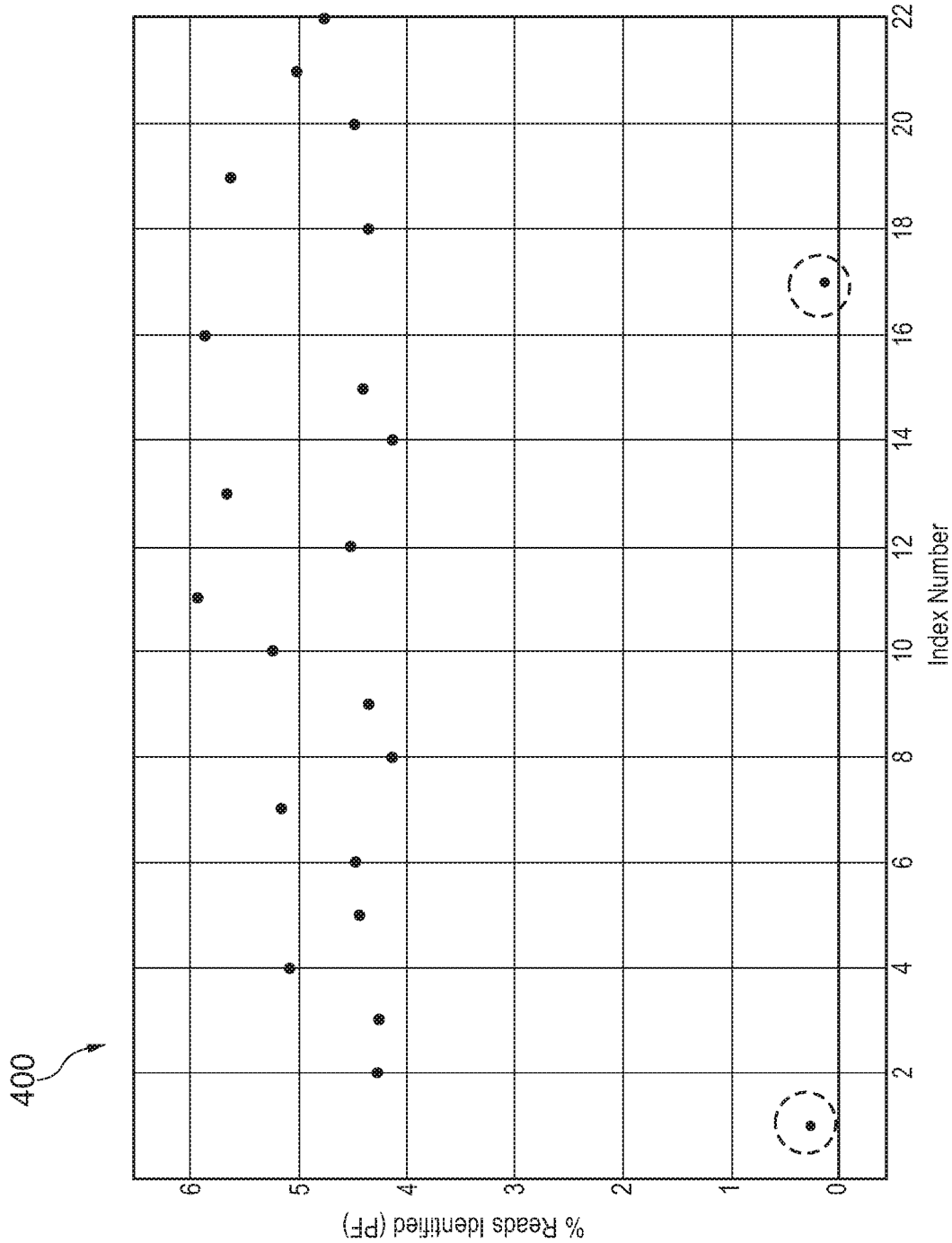


Fig. 4

5/16

500

| dbSNP ID   | CS219 |      | CS220 |      | CS221 |      | CS222 |      | CS223 |      | CS224 |      | CS225 |      |
|------------|-------|------|-------|------|-------|------|-------|------|-------|------|-------|------|-------|------|
|            | gDNA  | spot | gDNA  | spot | gDNA  | spot | gDNA  | spot | gDNA  | spot | gDNA  | spot | gDNA  | spot |
| rs2229546  | alt   | alt  | het   | het  | ref   | ref  | alt   | alt  | het   | het  | het   | het  | het   | het  |
| rs1410592  | ref   | ref  | ref   | ref  | alt   | alt  | alt   | alt  | het   | het  | het   | het  | het   | het  |
| rs497692   | alt   | alt  | het   | het  | alt   | alt  | het   | het  | alt   | alt  | alt   | alt  | het   | het  |
| rs10203363 | ref   | ref  | ref   | ref  | het   | het  | het   | het  | het   | het  | het   | het  | ref   | ref  |
| rs2819561  | alt   | alt  | alt   | alt  | het   | het  | alt   | alt  | het   | het  | het   | het  | alt   | alt  |
| rs4688963  | ref   | ref  | ref   | ref  | het   | het  | het   | het  | het   | het  | het   | het  | het   | het  |
| rs309557   | het   | het  | alt   | alt  | het   | het  | ref   | ref  | het   | het  | alt   | alt  | het   | het  |
| rs2942     | ref   | ref  | alt   | alt  | ref   | ref  | alt   | alt  | het   | het  | het   | het  | het   | het  |
| rs17548783 | ref   | ref  | ref   | ref  | het   | het  | alt   | alt  | het   | het  | alt   | alt  | het   | het  |
| rs4735258  | het   | het  | ref   | ref  | alt   | alt  | het   | het  | het   | het  | het   | het  | ref   | ref  |
| rs1381532  | alt   | alt  | het   | het  | alt   | alt  | ref   | ref  | het   | het  | ref   | ref  | ref   | ref  |
| rs10883099 | het   | het  | ref   | ref  | het   | het  | ref   | ref  | het   | het  | ref   | ref  | het   | het  |
| rs4617548  | het   | het  | ref   | ref  | het   | het  | het   | het  | ref   | ref  | alt   | alt  | het   | het  |
| rs7300444  | ref   | ref  | alt   | alt  | het   | het  | ref   | ref  | het   | het  | ref   | ref  | het   | het  |
| rs9532292  | alt   | alt  | het   | het  | ref   | ref  | alt   | alt  | het   | het  | ref   | ref  | het   | het  |
| rs2297995  | ref   | ref  | het   | het  | het   | het  | het   | het  | het   | het  | alt   | alt  | alt   | alt  |
| rs4577050  | alt   | alt  | het   | het  | alt   | alt  | het   | het  | alt   | alt  | het   | het  | alt   | alt  |
| rs2070203  | ref   | ref  | ref   | ref  | alt   | alt  | ref   | ref  | het   | het  | alt   | alt  | het   | het  |
| rs1037256  | alt   | alt  | alt   | alt  | ref   | ref  | het   | het  | het   | het  | ref   | ref  | ref   | ref  |
| rs9962023  | het   | het  | alt   | alt  | alt   | alt  | het   | het  | het   | het  | alt   | alt  | het   | het  |
| rs2228611  | alt   | alt  | het   | het  | alt   | alt  | ref   | ref  | ref   | ref  | het   | het  | het   | het  |
| rs10373    | ref   | ref  | het   | het  | het   | het  | het   | het  | alt   | alt  | ref   | ref  | het   | het  |
| rs4148973  | het   | het  | alt   | alt  | alt   | alt  | het   | het  | het   | het  | ref   | ref  | het   | het  |
| rs4148973  | ref   | ref  | het   | het  | ref   | ref  | ref   | ref  | het   | het  | het   | het  | alt   | alt  |

Fig. 5

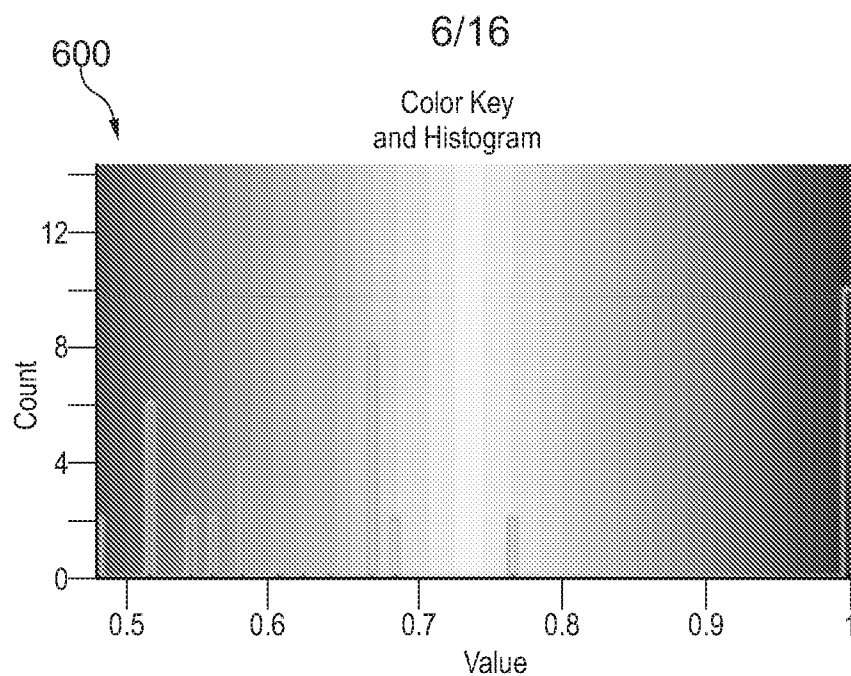


Fig. 6A

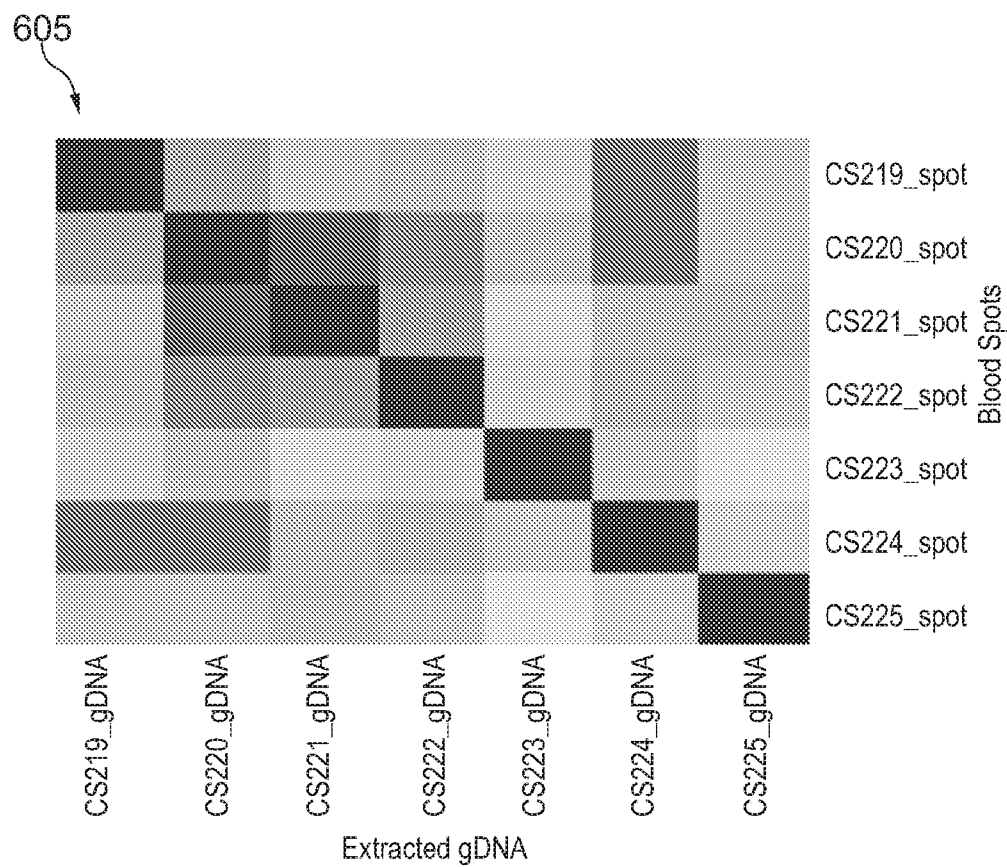


Fig. 6B

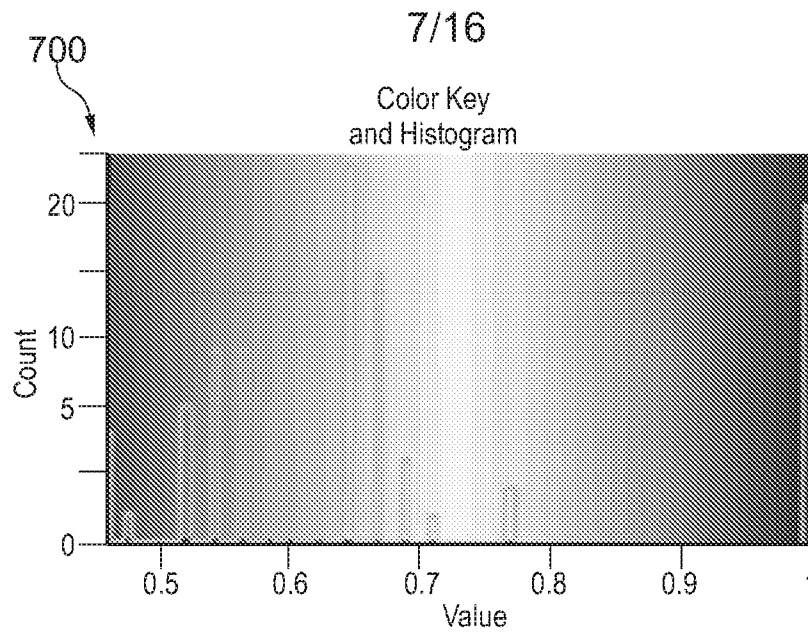


Fig. 7A

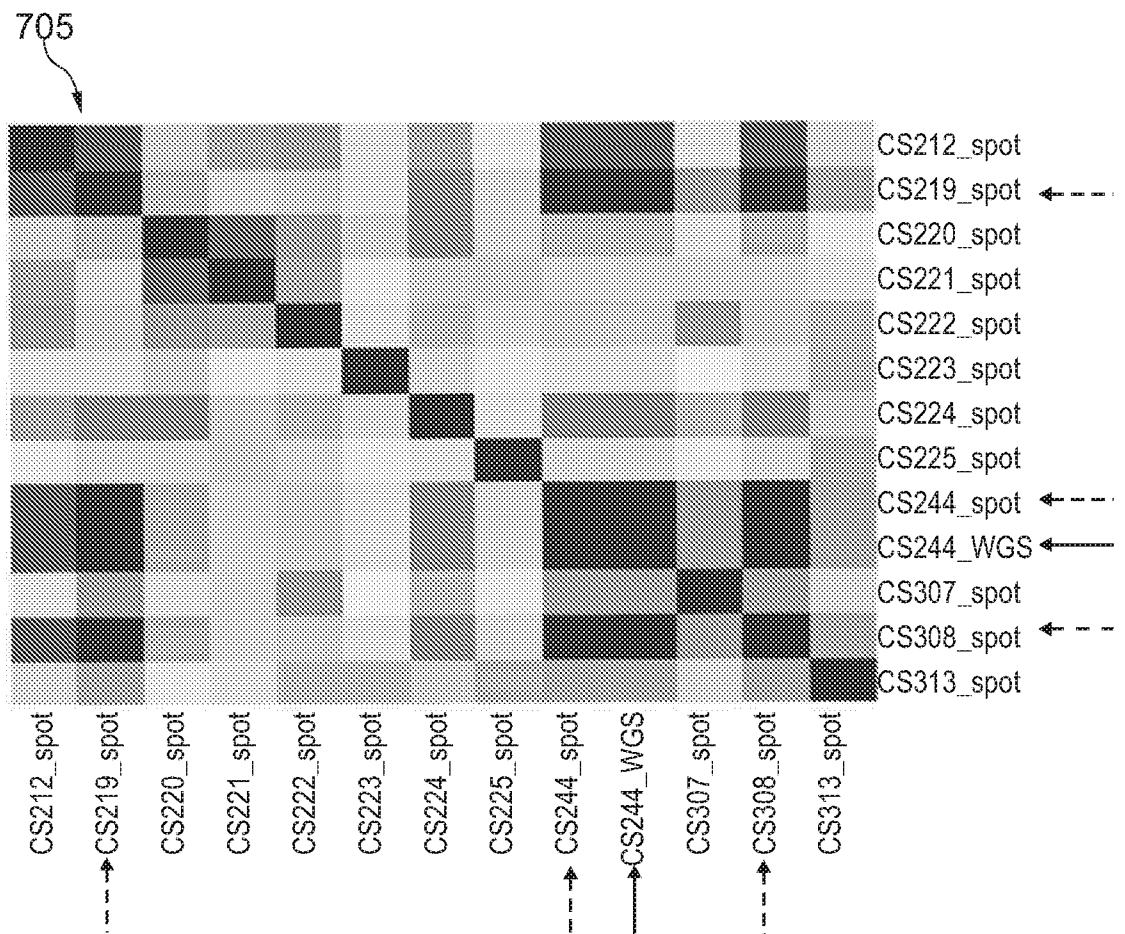


Fig. 7B

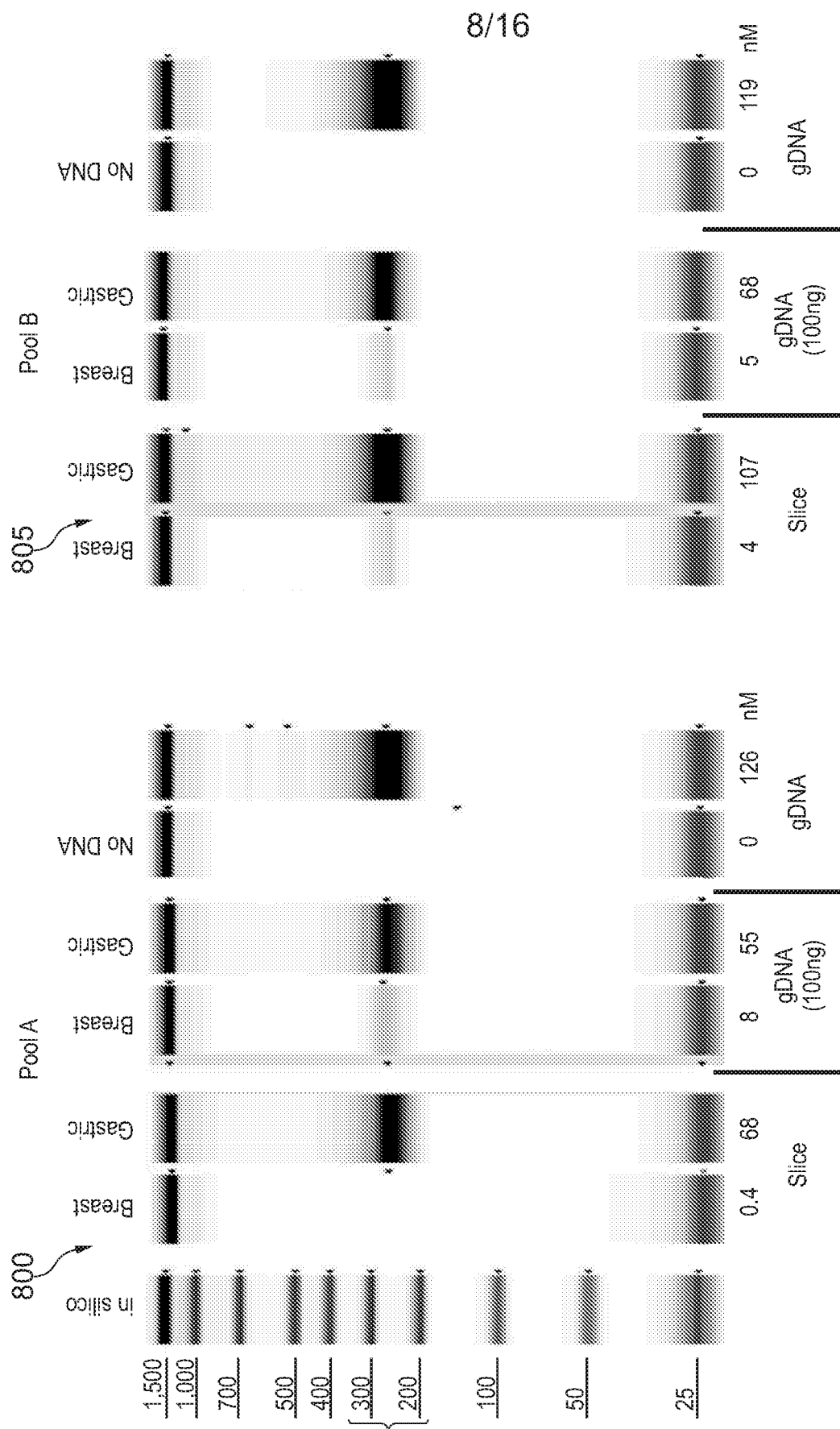


Fig. 8

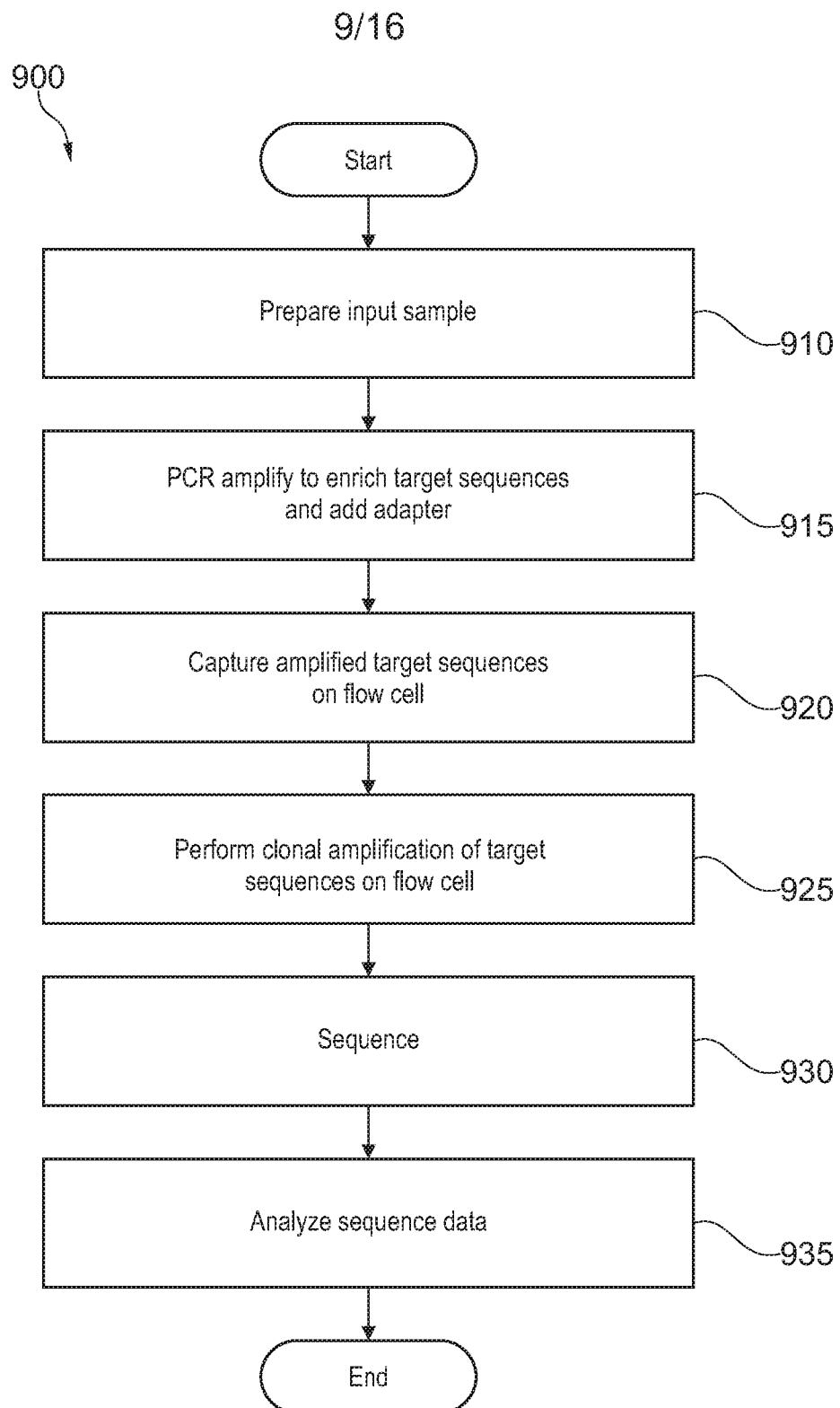


Fig. 9

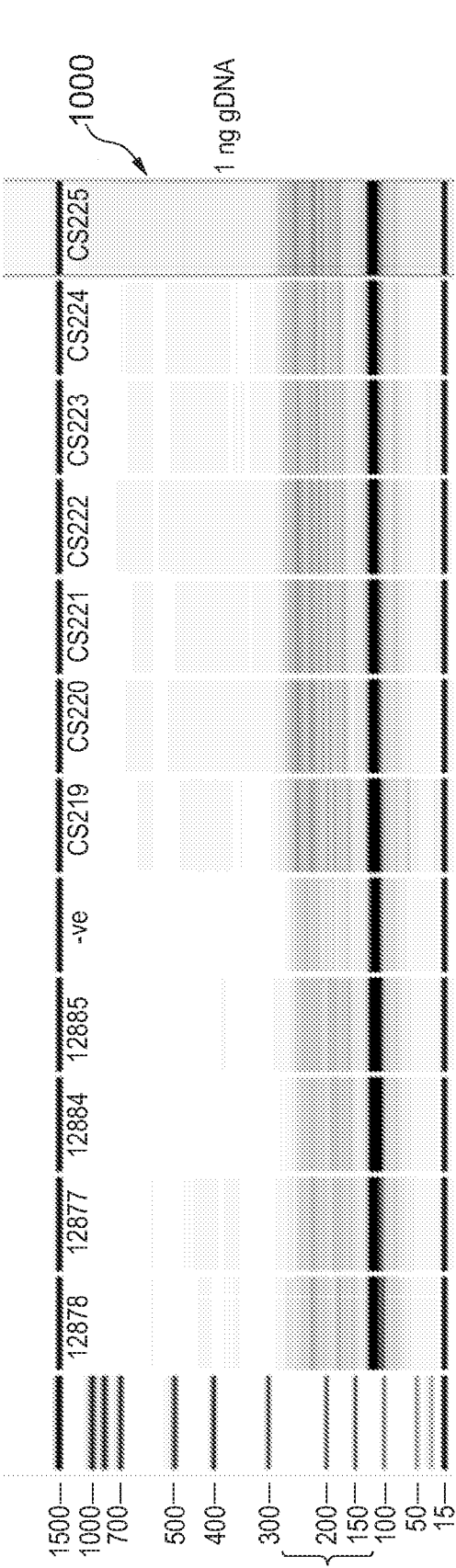


Fig. 10A

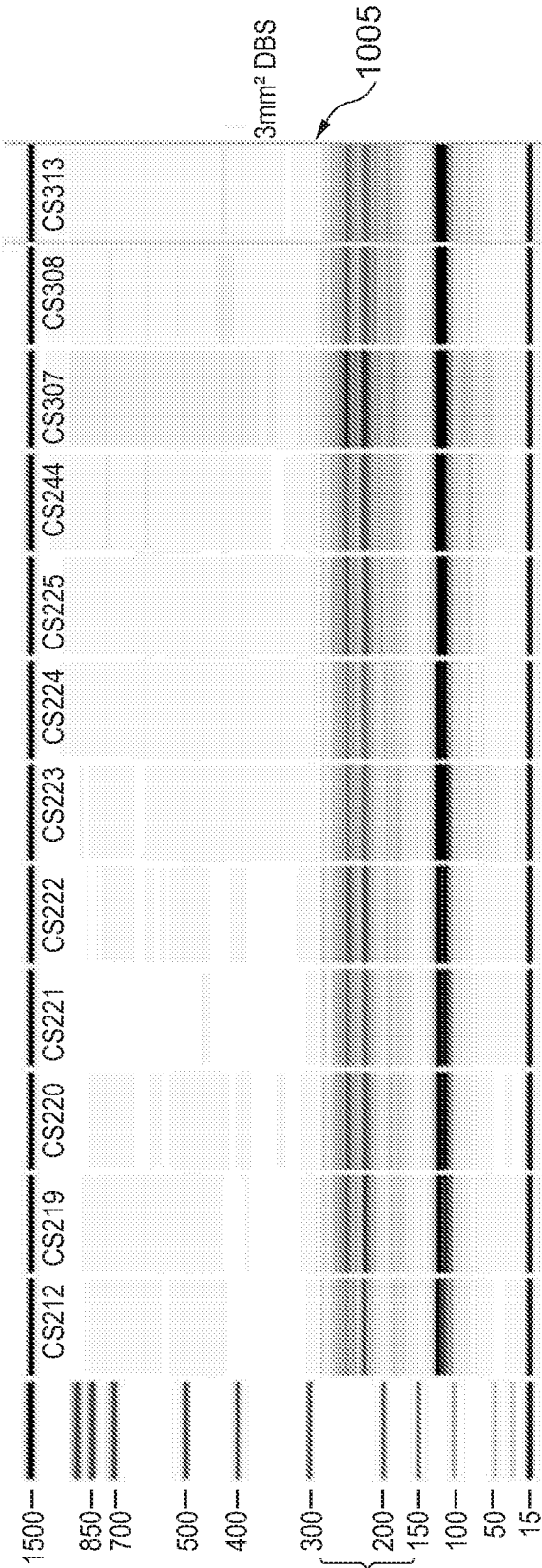


Fig. 10B

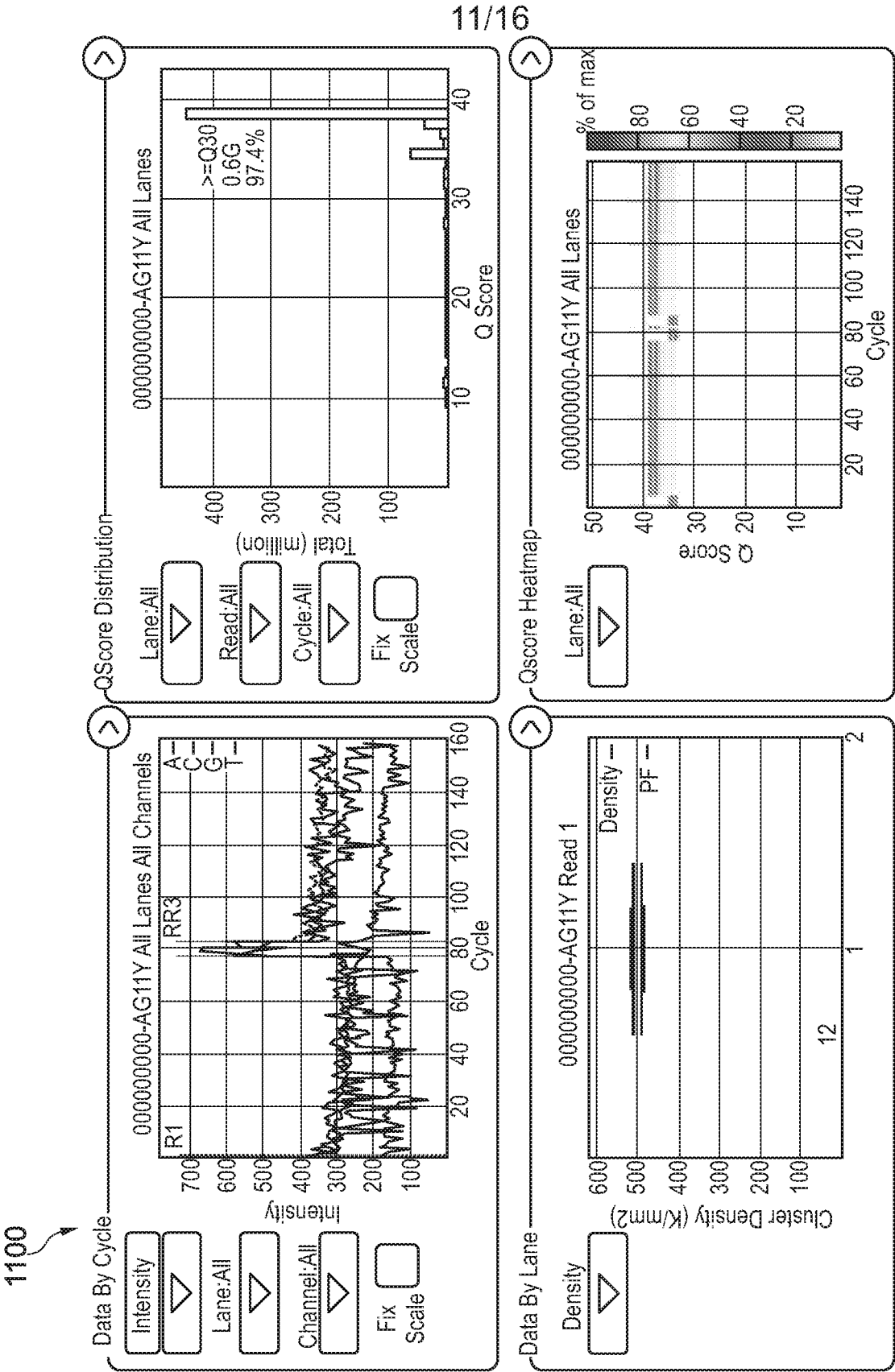


Fig. 11

12/16

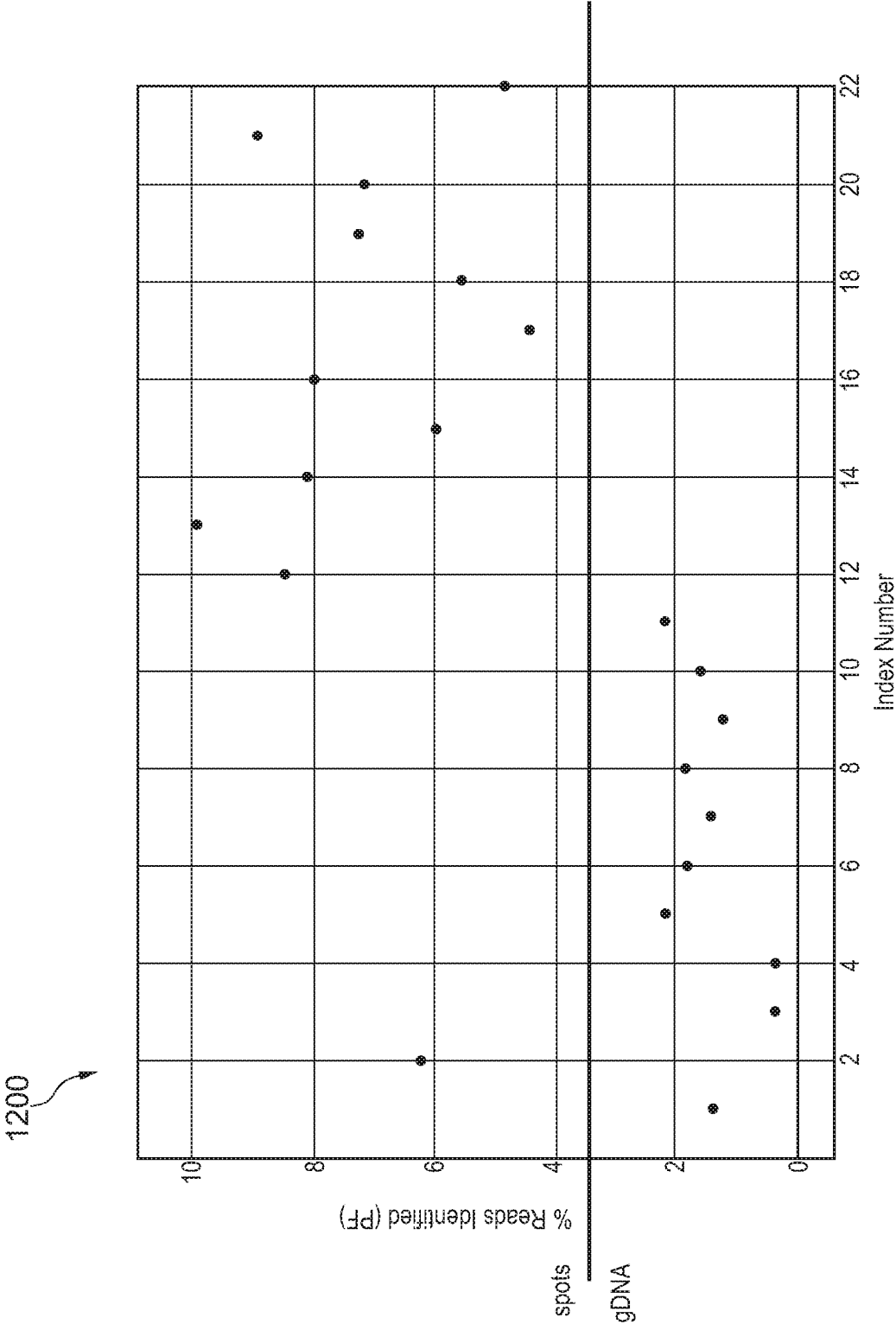


Fig. 12

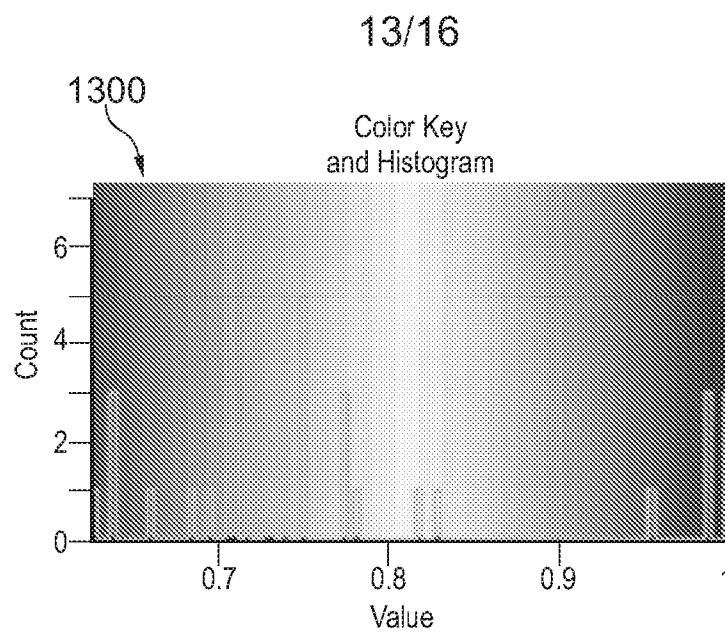


Fig. 13A

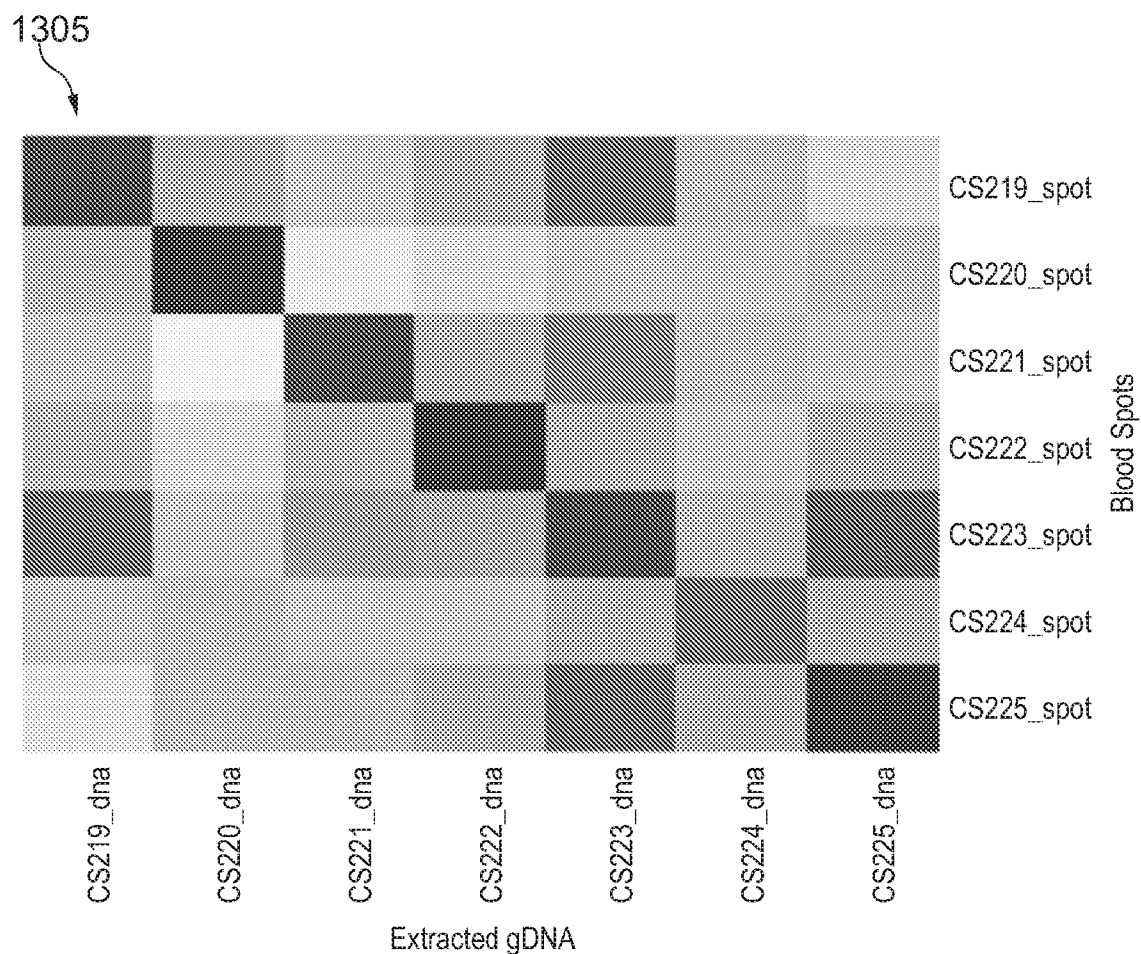


Fig. 13B

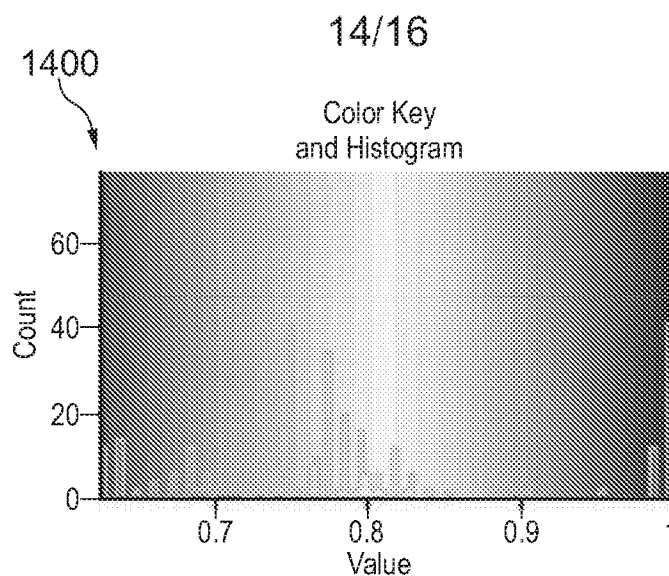


Fig. 14A

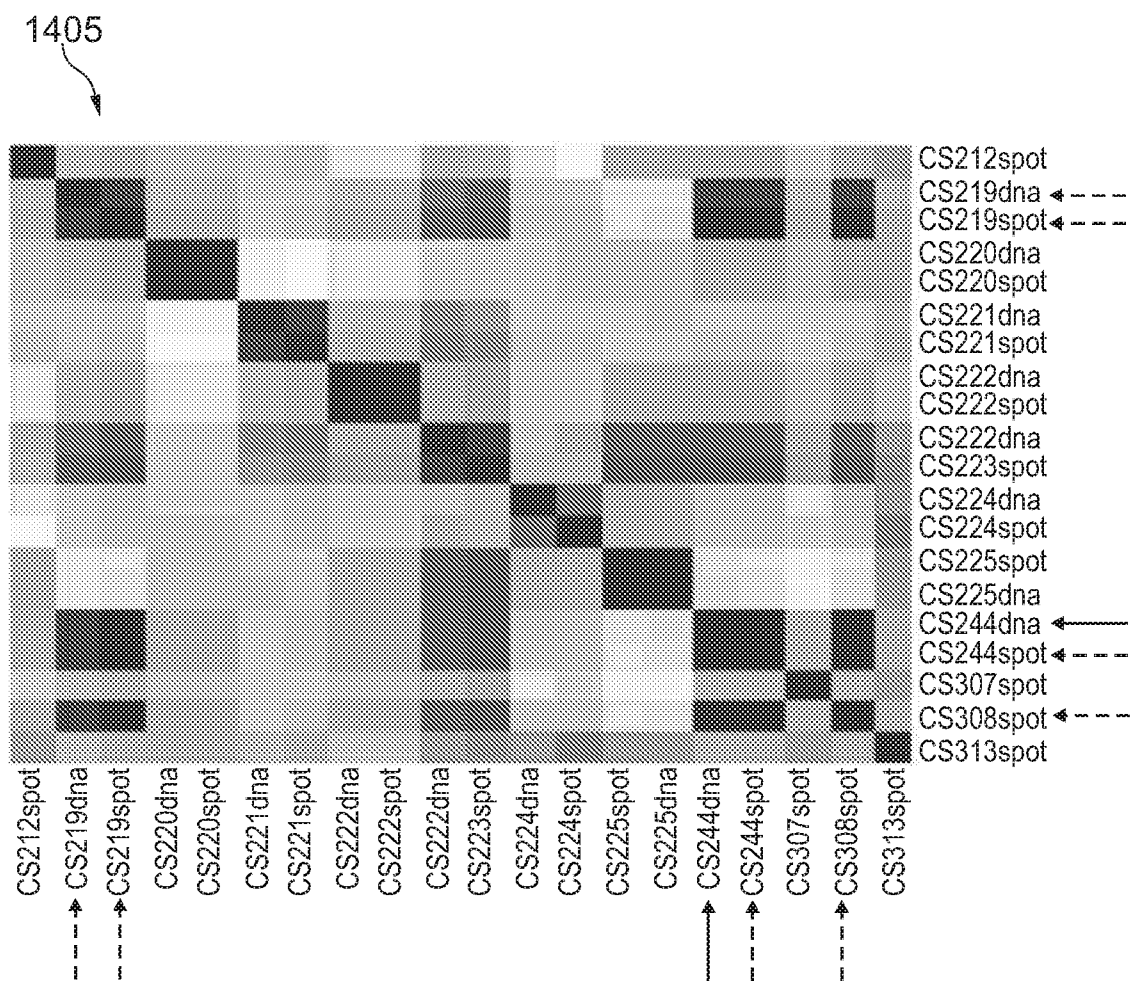


Fig. 14B

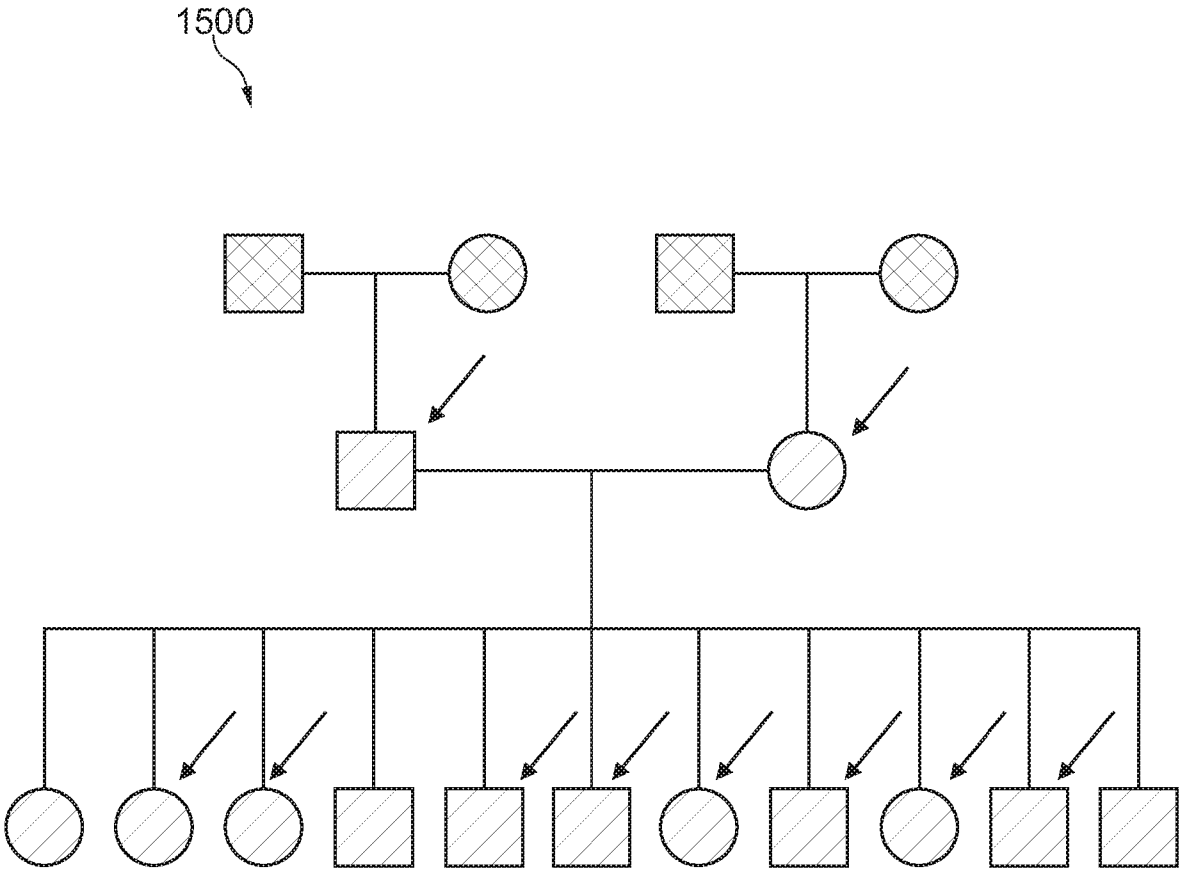


Fig. 15

16/16

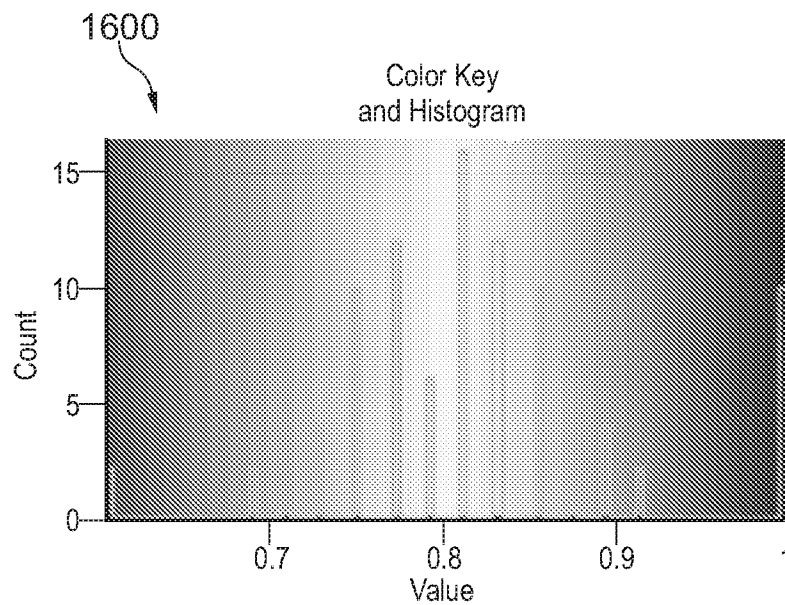


Fig. 16A

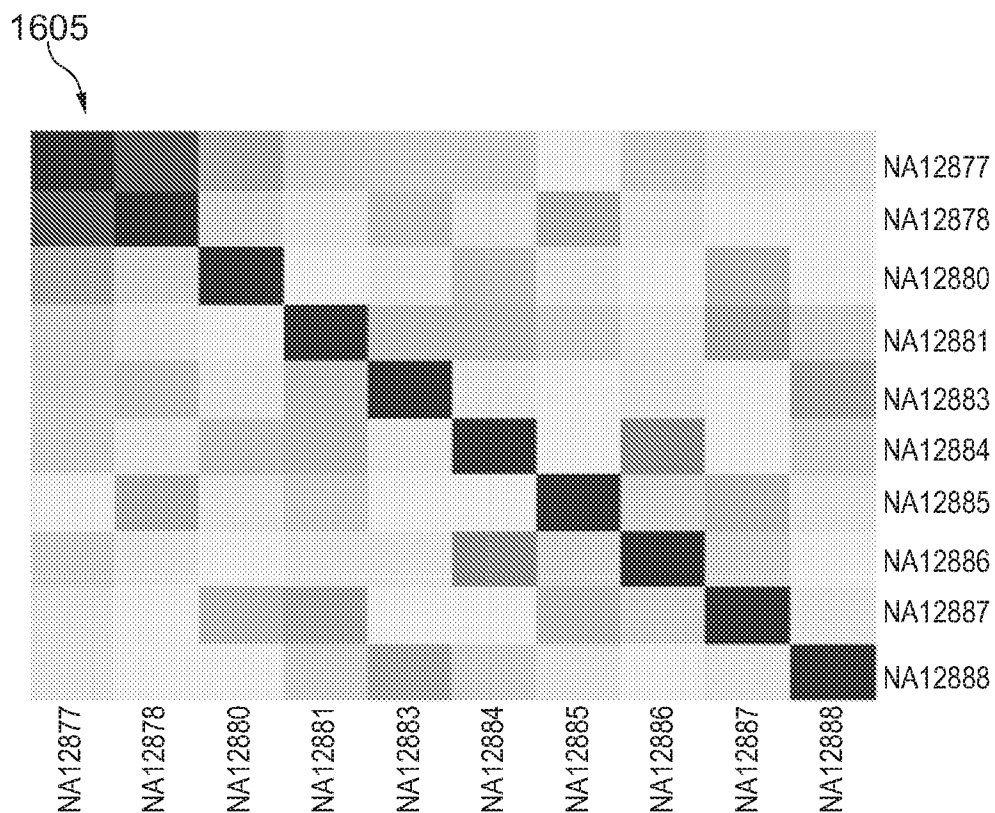


Fig. 16B

## INTERNATIONAL SEARCH REPORT

International application No  
PCT/GB2016/052026

A. CLASSIFICATION OF SUBJECT MATTER  
INV. C12Q1/68  
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)  
C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data, BIOSIS, EMBASE

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                      | Relevant to claim No.                                          |
|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|
| X         | SU JEONG PARK ET AL: "Direct STR Amplification from Whole Blood and Blood- or Saliva-Spotted FTA without DNA Purification", JOURNAL OF FORENSIC SCIENCES, vol. 53, no. 2, 1 March 2008 (2008-03-01), pages 335-341, XP055072543, ISSN: 0022-1198, DOI: 10.1111/j.1556-4029.2008.00666.x | 31-33,<br>42,43,<br>45,<br>47-54,<br>58,60                     |
| Y         | page 335, column 2 - page 336, column 1, paragraph 1; figures 1-4<br>the whole document<br><br>-----<br>-/--                                                                                                                                                                            | 34-41,<br>44,46,<br>56,57,<br>59,<br>61-67,<br>69-77,<br>79,80 |



Further documents are listed in the continuation of Box C.



See patent family annex.

\* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

23 August 2016

Date of mailing of the international search report

06/09/2016

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040,  
Fax: (+31-70) 340-3016

Authorized officer

Schmitt, Anja

## INTERNATIONAL SEARCH REPORT

International application No

PCT/GB2016/052026

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                                                                                                                                                                        | Relevant to claim No.                                          |
|-----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|
| X         | ANASTASIA FEDICK ET AL: "High-throughput real-time PCR-based genotyping without DNA purification",<br>BMC RESEARCH NOTES, BIOMED CENTRAL LTD, GB,<br>vol. 5, no. 1,<br>19 October 2012 (2012-10-19), page 573,<br>XP021122874,<br>ISSN: 1756-0500, DOI:<br>10.1186/1756-0500-5-573                                                                                                                                                        | 31-33,<br>42,43,<br>45,47,<br>48,<br>51-53,<br>58,60           |
| Y         | page 2                                                                                                                                                                                                                                                                                                                                                                                                                                    | 34-41,<br>44,46,<br>56,57,<br>59,<br>61-67,<br>69-77,<br>79,80 |
| X         | -----<br>ELSA HERRAEZ-HERNANDEZ HERRAEZ-HERNANDEZ ET AL: "Detection and Genotyping of Human Papillomavirus DNA in Formalin- Fixed Paraffin-Embedded Specimens with the HPV Direct Flow CHIP System",<br>THE OPEN VIROLOGY JOURNAL,<br>vol. 7, no. 1,<br>18 October 2013 (2013-10-18), pages 91-95,<br>XP055296677,<br>NL<br>ISSN: 1874-3579, DOI:<br>10.2174/1874357920130927004                                                          | 51,52,<br>55,58,60                                             |
| Y         | page 92, column 1, paragraph 3 - column 2, paragraph 1                                                                                                                                                                                                                                                                                                                                                                                    | 1-16,<br>18-51,<br>53,54,<br>56,57,<br>59,61-80                |
| Y         | -----<br>YOUNG NYUN PARK ET AL: "Detection of Hepatitis C Virus RNA Using Ligation-Dependent Polymerase Chain Reaction in Formalin-Fixed, Paraffin-Embedded Liver Tissues",<br>AMERICAN JOURNAL OF PATHOLOGY; [10640],<br>ELSEVIER INC, US,<br>vol. 149, no. 5,<br>1 November 1996 (1996-11-01), pages 1485-1491, XP008008602,<br>ISSN: 0002-9440<br>page 1486, column 2, paragraph 2 - page 1488, column 1, paragraph 1<br>-----<br>-/-- | 1-16,<br>18-80                                                 |

## INTERNATIONAL SEARCH REPORT

International application No

PCT/GB2016/052026

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                                                               | Relevant to claim No.                                           |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|
| Y         | CRAIG APRIL ET AL: "Whole-Genome Gene Expression Profiling of Formalin-Fixed, Paraffin-Embedded Tissue Samples", PLOS ONE, vol. 4, no. 12, 3 December 2009 (2009-12-03), page e8162, XP055195128, DOI: 10.1371/journal.pone.0008162 page 9 the whole document                                                                    | 1-16,<br>18-80                                                  |
| Y         | US 2008/242555 A1 (SHEN MIN-JUI RICHARD [US] ET AL) 2 October 2008 (2008-10-02) claims 12-22 the whole document                                                                                                                                                                                                                  | 1-16,<br>18-80                                                  |
| Y         | ZHIAN ZHANG ET AL: "Direct DNA Amplification from Crude Clinical Samples Using a PCR Enhancer Cocktail and Novel Mutants of Taq", THE JOURNAL OF MOLECULAR DIAGNOSTICS, vol. 12, no. 2, 1 March 2010 (2010-03-01), pages 152-161, XP055001258, ISSN: 1525-1578, DOI: 10.2353/jmoldx.2010.090070 pages 153-154 the whole document | 31-54,<br>56-67,<br>69-77,<br>79,80                             |
| X,P       | WO 2015/189588 A1 (ILLUMINA CAMBRIDGE LTD [GB]) 17 December 2015 (2015-12-17) cited in the application                                                                                                                                                                                                                           | 1-3,<br>10-12,<br>15,16,<br>31-33,<br>40-42,<br>45-56,<br>58,60 |
| Y,P       | page 24, paragraph 4 - page 32, paragraph 3; claims 1-31; figures 8,12,15,21 the whole document                                                                                                                                                                                                                                  | 4-9,13,<br>14,<br>18-30,<br>34-39,<br>41-43,<br>57,59,<br>61-80 |

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/GB2016/052026

| Patent document<br>cited in search report | Publication<br>date | Patent family<br>member(s) | Publication<br>date |
|-------------------------------------------|---------------------|----------------------------|---------------------|
| US 2008242555 A1                          | 02-10-2008          | AU 2003272203 A1           | 06-01-2004          |
|                                           |                     | US 2003211489 A1           | 13-11-2003          |
|                                           |                     | US 2008242555 A1           | 02-10-2008          |
|                                           |                     | WO 2004001062 A2           | 31-12-2003          |
| -----                                     |                     |                            |                     |
| WO 2015189588 A1                          | 17-12-2015          | US 2015353989 A1           | 10-12-2015          |
|                                           |                     | WO 2015189588 A1           | 17-12-2015          |
| -----                                     |                     |                            |                     |