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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 March 2016 (17.03.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/040446 A1

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

(21) International Application Number:
PCT/US2015/049132

(22) International Filing Date:
9 September 2015 (09.09.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/048,452 10 September 2014 (10.09.2014) US

(71) Applicant: **GOOD START GENETICS, INC.** [US/US]; 237 Putnam Avenue, Cambridge, MA 02139 (US).

(72) Inventors: **GOLE, Jeff**; 237 Putnam Avenue, Cambridge, MA 02139 (US). **GORE, Athurva**; 237 Putnam Avenue, Cambridge, MA 02139 (US). **UMBARGER, Mark**; 5 Winchester Street, Unit 101, Brookline, MA 02446 (US).

(74) Agents: **MEYERS, Thomas, C.** et al.; Brown Rudnick LLP, One Financial Center, Boston, MA 02111 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, 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: METHODS FOR SELECTIVELY SUPPRESSING NON-TARGET SEQUENCES

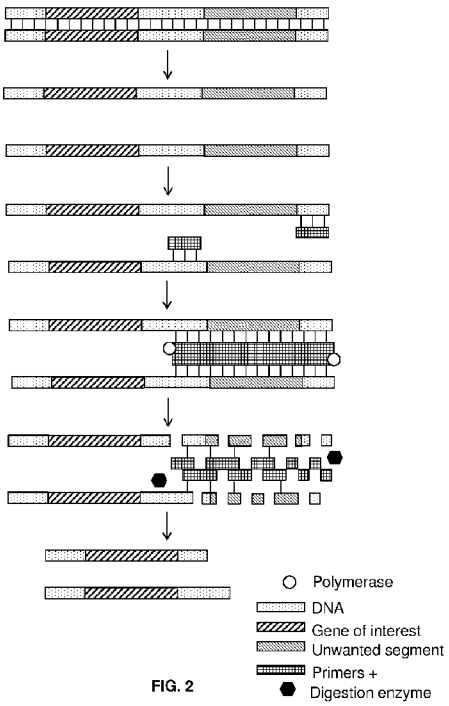

FIG. 2

(57) Abstract: The invention generally relates to negative selection of nucleic acids. The invention provides methods and systems that remove unwanted segments of nucleic acid in a sample so that a target gene or region of interest may be analyzed without interference from the unwanted segments. A sample is obtained that includes single-stranded nucleic acid with one or more unwanted segments. Complementary nucleic acid is added to the single-stranded nucleic acid to create a double-stranded region that includes the unwanted segment. The double-stranded region is then digested, leaving single-stranded nucleic acid that includes the target gene or region of interest. This allows paralogs, pseudogenes, repetitive elements, and other segments of the genome that may be similar to the target gene or region of interest to be removed from the sample.

WO 2016/040446 A1

METHODS FOR SELECTIVELY SUPPRESSING NON-TARGET SEQUENCES

Cross-Reference to Related Application

This application claims priority to, and the benefit of, U.S. Provisional Patent Application Serial No. 62/048,452, filed September 10, 2014, the contents of which are incorporated by reference.

Technical Field

The invention generally relates to negative selection of nucleic acids.

Background

The advent of high-throughput DNA sequencing has the potential to revolutionize modern biology and transform diagnostic medicine. Instruments for next-generation sequencing (NGS) continue to generate more data and become more inexpensive at a rate far outpacing Moore's Law. However, the most popular sequencers have an extremely short read length, limiting their ability to characterize any gene containing paralogous sequence or repetitive elements. As nearly two thirds of the genome is highly repetitive and over 20,000 pseudogenic regions exist, much of the genome is very difficult to characterize in a modern whole-genome sequencing experiment. Unfortunately, for many genes of clinical interest, characterizing those genes is made difficult by the presence of paralogs, pseudogenic homologs, and other segments of the genome that may be similar to the gene of interest and thus stymie attempts to detect, sequence, or isolate the gene of interest. As a result, despite the power of NGS instruments, some disease-related genes and mutations, even where known, are difficult to detect.

Summary

The invention provides methods and systems that remove unwanted segments of nucleic acid in a sample so that a target gene or region of interest may be analyzed without interference from the unwanted segments. A sample is obtained that includes single-stranded nucleic acid with one or more unwanted segments. Primers that are specific or preferentially bind to the unwanted segment are hybridized to the single-stranded nucleic acid within the unwanted region

or in a non-repetitive section upstream of the unwanted region and extended by a polymerase to create a double-stranded region that includes the unwanted segment. The double-stranded region is then digested, leaving single-stranded nucleic acid that includes the target gene or region of interest. This allows paralogs, pseudogenes, repetitive elements, and other segments of the genome that may be similar to the target gene or region of interest to be removed from the sample. The target gene or region of interest may thus be detected or characterized by analysis without interference from the unwanted segments. This may provide an improved ability to detect features such as disease-related genes and mutations, thus improving the clinical value of NGS technologies.

Systems and methods of the invention may be used to remove unwanted regions from genomic DNA (such as homologous genes, pseudogenes, or repetitive elements) prior to any DNA-based experimental procedure, including but not limited to microarray hybridization, quantitative or standard polymerase chain reaction, multiplex target capture, or DNA sequencing (either targeted or shotgun). Systems and methods of the invention provide for the identification of mutations in previously difficult-to-characterize genes, and therefore allow practitioners to expand the number of genes included in a targeted or whole-genome sequencing assay.

In certain aspects, the invention provides a method of removing unwanted segments of a nucleic acid from a sample. The method includes annealing a nucleic acid primer to a portion of a single-stranded nucleic acid that flanks an unwanted segment of the nucleic acid, extending the annealed primer in order to create a double-stranded region that includes the unwanted segment; and digesting the double-stranded region, thereby removing the unwanted segment from the nucleic acid.

The nucleic acid in the sample may include DNA, RNA, modified nucleic acids, or combinations thereof. The method may include obtaining a sample from a subject and denaturing double-stranded DNA in the sample. Denaturing can include the use of methods such as exposing the sample to heat, a detergent, or an acidic or basic solution.

The primer may be annealed within the unwanted segment or within an area upstream of the unwanted segment and extended. A pair or a number of primers may be used and primers that flank the unwanted segment may be used. In certain embodiments, a plurality of primers are annealed to a plurality of portions of that nucleic acid that flank an unwanted segment. The primer or primers are preferably extended using a polymerase enzyme under conditions

sufficient to cause extension of the primer in a template-dependent manner. In some embodiments, a primer or oligonucleotide is hybridized to the unwanted segment to create the double-stranded region containing the unwanted segment without need for an extension step.

The double-stranded region is digested. This can include exposing the sample to an enzyme that preferentially digests double-stranded nucleic acid such as certain double-stranded endonucleases, restriction endonucleases, or nicking enzymes. After digestion, the enzyme may be de-activated (e.g., by heat, chemicals, etc.). Digestion preferably results in intact genomic DNA lacking one or more unwanted segment and that is compatible with a nucleic acid analysis assay. Nucleic acid that is not digested may be analyzed by a nucleic acid analysis assay.

Assays suitable for analysis of the remaining un-digested nucleic acid may make use of molecular inversion probe capture, hybrid capture, Haloplex, sequencing (e.g., Sanger sequencing, NGS, or both), other methodologies, or combinations thereof. Where the unwanted segment is a paralog, a pseudogene, or non-paralogous repetitive element, such elements may be removed from the sample by methods of the invention.

In certain aspects, the invention provides a method of removing nucleic acid from a sample. The method includes annealing at least one oligonucleotide to single-stranded DNA in a sample, wherein the single-stranded DNA comprises target and non-target sequence. The oligonucleotide may be annealed to the non-target sequence to create double-stranded DNA that includes the non-target sequence or the oligonucleotide may be annealed elsewhere and extended to create double-stranded DNA that includes the non-target sequence. The non-target sequence is removed from the sample by digesting the double-stranded DNA. The target sequence may be analyzed using, e.g., molecular inversion probes, microarray hybridization, multiplex ligation-dependent probe amplification (MLPA), sequencing, fingerprinting techniques such as RFLP/ AFLP, chromatography, others, or combinations thereof. In some embodiments, the method includes first obtaining the sample from a subject and denaturing double-stranded subject DNA to produce the single-stranded DNA. Preferably, that single-stranded DNA consists essentially of genomic DNA from the subject prior to the annealing of the oligo. The annealing may include annealing a pair of oligonucleotides to the single-stranded DNA at sites that flank the non-target sequence (i.e., to remove both strands of the unwanted segment or non-target sequence. In certain embodiments, the target and non-target sequence are both located on at least one single strand of the single-stranded DNA, and extending the at least one oligonucleotide and digesting

the double-stranded DNA results in removing the non-target sequence from the at least one single strand of the single-stranded DNA.

Brief Description of the Drawings

FIG. 1 diagrams a method of removing unwanted segments of a nucleic acid.

FIG. 2 illustrates methods according to certain embodiments.

FIG. 3 gives a diagram of a system according to embodiments of the invention.

Detailed Description

To enable the characterization of difficult genomic regions using high-throughput short-read sequencing, the invention provides methods for the removal of unwanted genomic regions from a population of DNA molecules (e.g. genomic DNA). Most DNA-based techniques rely on the amplification of specific regions of interest or sequencing library molecules in a positive selection process (e.g. amplification utilizing primers that are unique to a single paralog). Methods of the invention instead involve a negative selection technique that removes any undesired analogous sequence, allowing application of standard high-throughput sequencing techniques or other analyses to any difficult-to-characterize gene of interest.

Applicability of methods of the invention may be illustrated by reference to two exemplary genes of interest for which direct high-throughput sequencing-based approaches are currently insufficient. One gene is "glucosidase beta acid," or GBA, which has been implicated as causative in Gaucher disease. Currently, long-range polymerase chain reaction experiments are required to characterize this gene, as a pseudogene with nearly identical sequence exists a mere 15,000 base pairs away. By removing this pseudogenic region using the invention, GBA can be characterized with high specificity, enabling construction of a genetic screen for Gaucher disease. This gene is a suitable target for methods of the invention, as it is relatively small and contains nearby unique flanking sequence.

An additional gene of interest is "survival of motor neuron 1," or SMN1. This gene has been implicated in spinal muscular atrophy. Currently, due to the presence of a paralogous gene known as SMN2 that is 100,000 base pairs away from SMN1, characterization of SMN1 is extremely challenging. By removing SMN2 using the invention, SMA could be screened for with a high-throughput sequencing approach that would not require a complex statistical model.

Additionally, novel causative mutations in genes such as SMN1 could also be identified. This gene is a suitable target for methods of the invention. It is of a suitable size and flanked by highly repetitive regions.

FIG. 1 diagrams a method 101 of removing unwanted segments of a nucleic acid from a sample according to embodiments of the invention. The method includes obtaining 105 a sample that includes nucleic acid. An oligonucleotide is annealed 109 to an unwanted segment of the nucleic acid or a portion of the nucleic acid that flanks an unwanted segment of the nucleic acid. In embodiments in which the oligonucleotide flanks the unwanted segment, the oligonucleotide is extended 113 to create a double-stranded region that includes the unwanted segment. In embodiments in which the oligonucleotide is annealed to the unwanted segment, a double-stranded region that includes the unwanted segment is created by virtue of the hybridization of the oligonucleotide at that segment. The double-stranded region is digested 117, thus removing the unwanted segment from the nucleic acid. This allows for a region or gene of interest to be analyzed 121.

The sample that includes nucleic acid may be obtained 105 by any suitable method. The sample may be obtained from a tissue or body fluid that is obtained in any clinically acceptable manner. Body fluids may include mucous, blood, plasma, serum, serum derivatives, bile, blood, maternal blood, phlegm, saliva, sweat, amniotic fluid, menstrual fluid, mammary fluid, follicular fluid of the ovary, fallopian tube fluid, peritoneal fluid, urine, and cerebrospinal fluid (CSF), such as lumbar or ventricular CSF. A sample may also be a fine needle aspirate or biopsied tissue. A sample also may be media containing cells or biological material. Samples may also be obtained from the environment (e.g., air, agricultural, water and soil) or may include research samples (e.g., products of a nucleic acid amplification reaction, or purified genomic DNA, RNA, proteins, etc.).

Isolation, extraction or derivation of genomic nucleic acids may be performed by methods known in the art. Isolating nucleic acid from a biological sample generally includes treating a biological sample in such a manner that genomic nucleic acids present in the sample are extracted and made available for analysis. Generally, nucleic acids are extracted using techniques such as those described in Green & Sambrook, 2012, Molecular Cloning: A Laboratory Manual 4 edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (2028 pages), the contents of which are incorporated by reference herein. A kit may be used to

extract DNA from tissues and bodily fluids and certain such kits are commercially available from, for example, BD Biosciences Clontech (Palo Alto, CA), Epicentre Technologies (Madison, WI), Gentra Systems, Inc. (Minneapolis, MN), and Qiagen Inc. (Valencia, CA). User guides that describe protocols are usually included in such kits.

It may be preferable to lyse cells to isolate genomic nucleic acid. Cellular extracts can be subjected to other steps to drive nucleic acid isolation toward completion by, e.g., differential precipitation, column chromatography, extraction with organic solvents, filtration, centrifugation, others, or any combination thereof. The genomic nucleic acid may be resuspended in a solution or buffer such as water, Tris buffers, or other buffers. In certain embodiments the genomic nucleic acid can be re-suspended in Qiagen DNA hydration solution, or other Tris-based buffer of a pH of around 7.5.

Any nucleic acid may be analyzed using methods of the invention. Nucleic acids suitable for use in aspects of the invention may include without limit genomic DNA, genomic RNA, synthesized nucleic acids, whole or partial genome amplification product, and high molecular weight nucleic acids, e.g. individual chromosomes. In certain embodiments, a sample is obtained that includes double-stranded DNA, such as bulk genomic DNA from a subject, and the double-stranded DNA is then denatured.

Double stranded nucleic acid may be denatured using any suitable method such as, for example, through the use of heat, detergent incubation, or an acidic or basic solution.

FIG. 2 illustrates the progress of methods according to certain embodiments. As shown in FIG. 2, methods may start with double stranded DNA (dotted shading if not otherwise hatched) that contains a gene of interest (first angled hatching pattern) and a paralog of the gene of interest (second angled hatching pattern). It will be appreciated that methods of the invention may operate starting with any suitable nucleic acid such as double- or single-stranded DNA or RNA or any combination thereof. The unwanted segment may be any sequence for which removal is desired from the starting nucleic acid. For example, the unwanted segment may include a paralog or homolog of a gene or region of interest; a pseudogene; or non-paralogous repetitive element. As used herein, homolog refers to a gene related to a second gene by descent from a common ancestral DNA sequence. Homolog describes the relationship between genes separated by the event of speciation (i.e., orthology) or to the relationship between genes separated by the event of genetic duplication (i.e., paralogy). Orthologs generally refers to genes in different species that

evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution and paralogs are genes related by duplication within a genome. See Fitch, 1970, Distinguishing homologs from analogous proteins, Syst Biol 19(2):99-113 and Jensen, 2001, Orthologs and paralogs—we need to get it right, Genome Biol 2(8):1002-1002.3. Pseudogenes include dysfunctional relatives of genes that have lost their protein-coding ability or are otherwise no longer expressed in the cell. Methods of the invention may be used to target a pseudogene that is present as a homolog to another gene or pseudogene within a sample and methods of the invention may be used to target a pseudogene that is present even where no known homologs of the pseudogene are suspected to also be present in the sample.

As illustrated in FIG. 2, the double-stranded DNA is denatured into its two complementary strands prior to primer hybridization. Any suitable method may be used to denature nucleic acid. Heat-based denaturing is a process by which double-stranded nucleic acid unwinds and separates into single-stranded strands. Heat denaturation of a nucleic acid of an unknown sequence typically uses a temperature high enough to ensure denaturation of even nucleic acids having a very high GC content, e.g., 95° C-98° C in the absence of any chemical denaturant. It is well within the abilities of one of ordinary skill in the art to optimize the conditions (e.g., time, temperature, etc.) for denaturation of the nucleic acid. Temperatures significantly lower than 95° C can also be used if the DNA contains nicks (and therefore sticky overhangs of low Tm), sequence of sufficiently low Tm, or chemical additives such as betaine.

Denaturing nucleic acids with the use of pH is also well known in the art, and such denaturation can be accomplished using any method known in the art such as introducing a nucleic acid to high or low pH, low ionic strength, and/or heat, which disrupts base-pairing causing a double-stranded helix to dissociate into single strands. For methods of pH-based denaturation see, for example, Ageno et al., 1969, The alkaline denaturation of DNA, Biophys J 9:1281-1311.

Nucleic acids can also be denatured via electro-chemical means, for example, by applying a voltage to a nucleic acid within a solution by means of an electrode. Varying methods of denaturing by applying a voltage are discussed in detail in U.S. Patent Nos. 6,197,508 and U.S. Patent No. 5,993,611. After denaturation, unwanted segments can be targeted for removal.

Methods of the invention include targeting unwanted segments of nucleic acid for removal. An unwanted segment of nucleic acid can be targeted for removal by making it into a

double-stranded segment. The unwanted segment can be made double-stranded by hybridizing a complementary oligonucleotide to the unwanted segment, by hybridizing a complementary oligonucleotide to a genomic segment flanking the unwanted segment and extending the oligonucleotide, or a combination thereof (e.g., an oligonucleotide can be hybridized so that it sits partially within the unwanted segment and then extended via methods described herein).

In certain embodiments, the oligonucleotide to be hybridized is a primer that is unique to the unwanted segment. For example, methods may include using a primer that is unique to a certain paralog or other element. The invention provides methods of making a primer and primer extension reactions that are unique to a paralog or similar segment by including or using a primer with a 3' end that terminates on a differentiating base (i.e., the 3'-most base or bases of the primer may be complementary to a base or bases that appear only in association with the segment (e.g., paralog) targeted for removal.

In some embodiments, double stranded DNA is created by hybridization alone (e.g., rather than by using oligonucleotide primer with polymerase extension). One or more long segments of nucleic acid complementary to the unwanted segments could be used. For example, long segments of synthetic DNA could be used. The segments of complementary nucleic acid could have any suitable length such as, for example, tens of bases, hundreds of bases, length of an exon, length of a gene, etc. Use of one or more long segments of nucleic acid complementary to the unwanted segments (e.g., followed by digestion of dsDNA) may provide for enrichment of, for example, target relative to non-target.

As noted above, the recognition site for the oligonucleotide, primer, or complementary nucleic acid may flank the unwanted segment, lie within the unwanted segment, or both. Additionally, methods may include using one or any suitable number of oligonucleotides or primers to target an unwanted segment or segments of nucleic acid.

In the non-limiting, illustrative embodiment shown in FIG. 2, primers (cross-hatching pattern) are annealed to unique genomic segments flanking the paralogous region. The primer may be annealed at any suitable location. For example, it may be preferable to anneal any of the one or more primers to a portion within 50 or fewer bases from the unwanted segment, although it may not be necessary to anneal the primers within 50 bases of the unwanted region. As shown in FIG. 2, primers are annealed at locations that flank the unwanted segment, i.e., each primer of a pair hybridizes to its target strand in a region that flanks the 5' end of the unwanted segment. In

this way, extension of the primers will result in most or all of the unwanted segments being present in exclusively double-stranded form, whereas the desired region(s) should remain in a primarily single-stranded state.

In certain embodiments, polymerase (drawn as an open circle in FIG. 2) is used to perform second-strand synthesis over the paralogous region. Extending the annealed primer creates a double-stranded region that includes the unwanted segment. The primer is extended using a polymerase enzyme under conditions sufficient to cause extension of the primer in a template-dependent manner. Suitable polymerase enzymes include phi29, Bst, Exo-minus E. Coli Polymerase I, Taq Polymerase, and T7 Polymerase I.

An enzymatic digestion (the digestion enzyme is represented by a darkened hexagon in FIG. 2) is then used to degrade only the double-stranded paralogous region, leaving behind the gene of interest. Any suitable digestion platform may be employed such as, for example, dsDNAse, fragmentase, a non-specific nicking enzyme such as a modified Vvn, restriction enzymes such as MspJI and FspEI, and a combination of USER plus T7 endonuclease I.

Thermo Scientific dsDNase is an engineered shrimp DNase designed for rapid and safe removal of contaminating genomic DNA from RNA samples. It is an endonuclease that cleaves phosphodiester bonds in DNA to yield oligonucleotides with 5'-phosphate and 3'-hydroxyl termini. Highly specific activity towards double-stranded DNA ensures that RNA and single-stranded DNA such as cDNA and primers are not cleaved. dsDNase is easily inactivated by moderate heat treatment (55°C). Thermo Scientific dsDNAse is available from Thermo Fisher Scientific, Inc. (Waltham, MA).

Fragmentase includes the enzyme sold under the trademark NEBNEXT dsDNA fragmentase by New England Biolabs (Ipswich, MA). NEBNEXT dsDNA fragmentase generates dsDNA breaks in a time-dependent manner to yield 50–1,000 bp DNA fragments depending on reaction time. NEBNext dsDNA Fragmentase contains two enzymes, one randomly generates nicks on dsDNA and the other recognizes the nicked site and cuts the opposite DNA strand across from the nick, producing dsDNA breaks. The resulting DNA fragments contain short overhangs, 5´-phosphates, and 3´-hydroxyl groups. The random nicking activity of NEBNext dsDNA Fragmentase has been confirmed by preparing libraries for next-generation sequencing. A comparison of the sequencing results between genomic DNA (gDNA) prepared with NEBNext dsDNA fragmentase and with mechanical shearing demonstrates that the NEBNext

dsDNA Fragmentase does not introduce any detectable bias during the sequencing library preparation and no difference in sequence coverage is observed using the two methods

The Vibrio vulnificus nuclease, Vvn, is a non-specific periplasmic nuclease capable of digesting DNA and RNA. It has been suggested that Vvn hydrolyzes DNA by a general single-metal ion mechanism. See Li, et al., 2003, DNA binding and cleavage by the periplasmic nuclease Vvn: a novel structure with a known active site, EMBO J 22(15):4014-4025.

MspJI is a modification dependent endonuclease that recognizes certain methylation patterns. The most common epigenetic modifications found in eukaryotic organisms are methylation marks at CpG or CHG sites. A subset of these modified sites are recognized and cleaved by MspJI. MspJI is available from New England Biolabs. T7 Endonuclease I recognizes and cleaves non-perfectly matched DNA, cruciform DNA structures, Holliday structures or junctions, hetero-duplex DNA and more slowly, nicked double-stranded DNA. The cleavage site is at the first, second or third phosphodiester bond that is 5´ to the mismatch. The protein is the product of T7 gene 3. Any other suitable enzyme for digesting the target unwanted segments may be used.

The added enzymes may then be deactivated using an irreversible heat or chemical treatment, leaving genomic DNA lacking an intact undesired region(s) yet still compatible with any downstream assay (e.g. molecular inversion probe capture or any other library construction methodology).

The digesting step results in intact genomic DNA lacking one or more unwanted segment and that is compatible with a nucleic acid analysis assay. This DNA can then be utilized for any downstream assay. Downstream assays may include molecular inversion capture, sequencing, others, or a combination thereof.

Methods of the invention can be used to negatively select out pseudogenic regions from the genome. Methods of the invention can be combined with a genetic test, screening, or other assay in order to screen patients for mutations in a gene (e.g., GBA, SMN1, or other genes containing paralogous regions). Some background may be found in published international patent application WO 2013/191775, to Nugen Technologies, Inc.

After removing the unwanted segment from the nucleic acid, the sample may be enriched for genes of interest using methods known in the art, such as hybrid capture. Methods suitable for use may be found discussed in U.S. Pat. 8,529,744; U.S. Pat. 7,985,716; U.S. Pat. 7,666,593;

and U.S. Pat. 6,613,516. As will be described in more detail below, a preferable capture method uses molecular inversion probes.

Nucleic acids, including genomic nucleic acids, can be fragmented using any of a variety of methods, such as mechanical fragmenting, chemical fragmenting, and enzymatic fragmenting. Methods of nucleic acid fragmentation are known in the art and include, but are not limited to, DNase digestion, sonication, mechanical shearing, and the like. U.S. Pub 2005/0112590 provides a general overview of various methods of fragmenting known in the art.

Genomic nucleic acids can be fragmented into uniform fragments or randomly fragmented. In certain aspects, nucleic acids are fragmented to form fragments having a fragment length of about 5 kilobases or 100 kilobases. Desired fragment length and ranges of fragment lengths can be adjusted depending on the type of nucleic acid targets one seeks to capture and the design and type of probes such as molecular inversion probes (MIPs) that will be used. Chemical fragmentation of genomic nucleic acids can be achieved using methods such as a hydrolysis reaction or by altering temperature or pH. Nucleic acid may be fragmented by heating a nucleic acid immersed in a buffer system at a certain temperature for a certain period to time to initiate hydrolysis and thus fragment the nucleic acid. The pH of the buffer system, duration of heating, and temperature can be varied to achieve a desired fragmentation of the nucleic acid. Mechanical shearing of nucleic acids into fragments can be used e.g., by hydro-shearing, trituration through a needle, and sonication. The nucleic acid can also be sheared via nebulization, hydro-shearing, sonication, or others. See U.S. Pat. 6,719,449; U.S. Pat. 6,948,843; and U.S. Pat. 6,235,501. Nucleic acid may be fragmented enzymatically. Enzymatic fragmenting, also known as enzymatic cleavage, cuts nucleic acids into fragments using enzymes, such as endonucleases, exonucleases, ribozymes, and DNAzymes. Varying enzymatic fragmenting techniques are well-known in the art. Additionally, DNA may be denatured again as needed after the digestion and any other sample prep steps. For example, during a fragmentation step, ssDNA may anneal to form dsDNA and it may be desirable to again denature the dsDNA. In certain embodiments, the sample nucleic acid is captured or targeted using any suitable capture method or assay such as hybridization capture or capture by probes such as MIPs.

MIPs, or molecular inversion probes, can be used to detect or amplify particular nucleic acid sequences in complex mixtures. Use of molecular inversion probes has been demonstrated for detection of single nucleotide polymorphisms (Hardenbol et al., 2005, Highly multiplexed

molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay, Genome Res 15:269-75) and for preparative amplification of large sets of exons (Porreca et al., 2007, Multiplex amplification of large sets of human exons, Nat Methods 4:931-6, Krishnakumar et al., 2008, A comprehensive assay for targeted multiplex amplification of human DNA sequences, PNAS 105:9296-301). One of the main benefits of the method is in its capacity for a high degree of multiplexing, because generally thousands of targets may be captured in a single reaction containing thousands of probes.

In certain embodiments, molecular inversion probes include a universal portion flanked by two unique targeting arms. The targeting arms are designed to hybridize immediately upstream and downstream of a specific target sequence located on a genomic nucleic acid fragment. The molecular inversion probes are introduced to nucleic acid fragments to perform capture of target sequences located on the fragments. According to the invention, fragmenting aids in capture of target nucleic acid by molecular inversion probes. As described in greater detail herein, after capture of the target sequence (e.g., locus) of interest, the captured target may further be subjected to an enzymatic gap-filling and ligation step, such that a copy of the target sequence is incorporated into a circle. Capture efficiency of the MIP to the target sequence on the nucleic acid fragment can be improved by lengthening the hybridization and gap-filing incubation periods. (See, e.g., Turner et al., 2009, Massively parallel exon capture and library-free resequencing across 16 genomes, Nature Methods 6:315-316.)

A library of molecular inversion probes may be created and used in capturing DNA of genomic regions of interests (e.g., SMN1, SMN2, control DNA). The library includes a plurality of oligonucleotide probes capable of capturing one or more genomic regions of interest (e.g., SMN1, SMN2 and control loci) within the samples to be tested.

The result of MIP capture as described above is a library of circular target probes, which then can be processed in a variety of ways. Adaptors for sequencing may be attached during common linker-mediated PCR, resulting in a library with non-random, fixed starting points for sequencing. For preparation of a shotgun library, a common linker-mediated PCR is performed on the circle target probes, and the post-capture amplicons are linearly concatenated, sheared, and attached to adaptors for sequencing. Methods for shearing the linear concatenated captured targets can include any of the methods disclosed for fragmenting nucleic acids discussed above.

In certain aspects, performing a hydrolysis reaction on the captured amplicons in the presence of heat is the desired method of shearing for library production.

In some embodiments, the amount of target nucleic acid and probe used for each reaction is normalized to avoid any observed differences being caused by differences in concentrations or ratios. In some embodiments, in order to normalize genomic DNA and probe, the genomic DNA concentration is read using a standard spectrophotometer or by fluorescence (e.g., using a fluorescent intercalating dye). The probe concentration may be determined experimentally or using information specified by the probe manufacturer.

Similarly, once a locus has been captured, it may be amplified and/or sequenced in a reaction involving one or more primers. The amount of primer added for each reaction can range from 0.1 pmol to 1 nmol, 0.15 pmol to 1.5 nmol (for example around 1.5 pmol). However, other amounts (e.g., lower, higher, or intermediate amounts) may be used.

A targeting arm may be designed to hybridize (e.g., be complementary) to either strand of a genetic locus of interest if the nucleic acid being analyzed is DNA (e.g., genomic DNA). For MIP probes, whichever strand is selected for one targeting arm will be used for the other one. In the context of RNA analysis, a targeting arm should be designed to hybridize to the transcribed RNA. It also should be appreciated that MIP probes referred to herein as "capturing" a target sequence are actually capturing it by template-based synthesis rather than by capturing the actual target molecule (other than for example in the initial stage when the arms hybridize to it or in the sense that the target molecule can remain bound to the extended MIP product until it is denatured or otherwise removed).

A targeting arm may include a sequence that is complementary to one allele or mutation (e.g., a SNP or other polymorphism, a mutation, etc.) so that the probe will preferentially hybridize (and capture) target nucleic acids having that allele or mutation. Sequence tags (also referred to as barcodes) may be designed to be unique in that they do not appear at other positions within a probe or a family of probes and they also do not appear within the sequences being targeted. Uniformity and reproducibility can be increased by designing multiple probes per target, such that each base in the target is captured by more than one probe.

The length of a capture molecule on a nucleic acid fragment (e.g., a target nucleic acid or sub-region thereof) may be selected based upon multiple considerations. For example, where analysis of a target involves sequencing, e.g., with a next-generation sequencer, the target length

should typically match the sequencing read-length so that shotgun library construction is not necessary. However, it should be appreciated that captured nucleic acids may be sequenced using any suitable sequencing technique as aspects of the invention are not limited in this respect.

It is also to be appreciated that some target nucleic acids on a nucleic acid fragment are too large to be captured with one probe. Consequently, it may be helpful to capture multiple sub-regions of a target nucleic acid in order to analyze the full target.

Methods of the invention also provide for combining the method of fragmenting the nucleic acid prior to capture with other MIP capture techniques that are designed to increase target uniformity, reproducibility, and specificity. Other MIP capture techniques are shown in U.S. Pub. 2012/0165202, incorporated by reference.

Multiple probes, e.g., MIPs, can be used to amplify each target nucleic acid. In some embodiments, the set of probes for a given target can be designed to 'tile' across the target, capturing the target as a series of shorter sub targets. In some embodiments, where a set of probes for a given target is designed to 'tile' across the target, some probes in the set capture flanking non-target sequence). Alternately, the set can be designed to 'stagger' the exact positions of the hybridization regions flanking the target, capturing the full target (and in some cases capturing flanking non-target sequence) with multiple probes having different targeting arms, obviating the need for tiling. The particular approach chosen will depend on the nature of the target set. For example, if small regions are to be captured, a staggered-end approach might be appropriate, whereas if longer regions are desired, tiling might be chosen. In all cases, the amount of bias-tolerance for probes targeting pathological loci can be adjusted by changing the number of different MIPs used to capture a given molecule.

Probes for MIP capture reactions may be synthesized on programmable microarrays because of the large number of sequences required. Because of the low synthesis yields of these methods, a subsequent amplification step is required to produce sufficient probe for the MIP amplification reaction. The combination of multiplex oligonucleotide synthesis and pooled amplification results in uneven synthesis error rates and representational biases. By synthesizing multiple probes for each target, variation from these sources may be averaged out because not all probes for a given target will have the same error rates and biases.

Using methods described herein, a single copy of a specific target nucleic acid may be amplified to a level that can be sequenced. Further, the amplified segments created by an

amplification process such as PCR may be, themselves, efficient templates for subsequent PCR amplifications.

Amplification or sequencing adapters or barcodes, or a combination thereof, may be attached to the fragmented nucleic acid. Such molecules may be commercially obtained, such as from Integrated DNA Technologies (Coralville, IA). In certain embodiments, such sequences are attached to the template nucleic acid molecule with an enzyme such as a ligase. Suitable ligases include T4 DNA ligase and T4 RNA ligase, available commercially from New England Biolabs (Ipswich, MA). The ligation may be blunt ended or via use of complementary overhanging ends. In certain embodiments, following fragmentation, the ends of the fragments may be repaired, trimmed (e.g. using an exonuclease), or filled (e.g., using a polymerase and dNTPs) to form blunt ends. In some embodiments, end repair is performed to generate blunt end 5' phosphorylated nucleic acid ends using commercial kits, such as those available from Epicentre Biotechnologies (Madison, WI). Upon generating blunt ends, the ends may be treated with a polymerase and dATP to form a template independent addition to the 3'-end and the 5'-end of the fragments, thus producing a single A overhanging. This single A can guide ligation of fragments with a single T overhanging from the 5'-end in a method referred to as T-A cloning. Alternatively, because the possible combination of overhangs left by the restriction enzymes are known after a restriction digestion, the ends may be left as-is, i.e., ragged ends. In certain embodiments double stranded oligonucleotides with complementary overhanging ends are used.

In certain embodiments, one or more bar code is attached to each, any, or all of the fragments. A bar code sequence generally includes certain features that make the sequence useful in sequencing reactions. The bar code sequences are designed such that each sequence is correlated to a particular portion of nucleic acid, allowing sequence reads to be correlated back to the portion from which they came. Methods of designing sets of bar code sequences is shown for example in U.S. Pat. 6,235,475, the contents of which are incorporated by reference herein in their entirety. In certain embodiments, the bar code sequences range from about 5 nucleotides to about 15 nucleotides. In a particular embodiment, the bar code sequences range from about 4 nucleotides to about 7 nucleotides. In certain embodiments, the bar code sequences are attached to the template nucleic acid molecule, e.g., with an enzyme. The enzyme may be a ligase or a polymerase, as discussed above. Attaching bar code sequences to nucleic acid templates is shown in U.S. Pub. 2008/0081330 and U.S. Pub. 2011/0301042, the content of each of which is

incorporated by reference herein in its entirety. Methods for designing sets of bar code sequences and other methods for attaching bar code sequences are shown in U.S. Pats. 6,138,077; 6,352,828; 5,636,400; 6,172,214; 6235,475; 7,393,665; 7,544,473; 5,846,719; 5,695,934; 5,604,097; 6,150,516; 7,537,897; 6172,218; and 5,863,722, the content of each of which is incorporated by reference herein in its entirety. After any processing steps (e.g., obtaining, isolating, fragmenting, amplification, or barcoding), nucleic acid can be sequenced.

Sequencing may be by any method known in the art. DNA sequencing techniques include classic dideoxy sequencing reactions (Sanger method) using labeled terminators or primers and gel separation in slab or capillary, sequencing by synthesis using reversibly terminated labeled nucleotides, pyrosequencing, 454 sequencing, Illumina/Solexa sequencing, allele specific hybridization to a library of labeled oligonucleotide probes, sequencing by synthesis using allele specific hybridization to a library of labeled clones that is followed by ligation, real time monitoring of the incorporation of labeled nucleotides during a polymerization step, polony sequencing, and SOLiD sequencing. Separated molecules may be sequenced by sequential or single extension reactions using polymerases or ligases as well as by single or sequential differential hybridizations with libraries of probes.

A sequencing technique that can be used includes, for example, Illumina sequencing. Illumina sequencing is based on the amplification of DNA on a solid surface using fold-back PCR and anchored primers. Genomic DNA is fragmented, and adapters are added to the 5' and 3' ends of the fragments. DNA fragments that are attached to the surface of flow cell channels are extended and bridge amplified. The fragments become double stranded, and the double stranded molecules are denatured. Multiple cycles of the solid-phase amplification followed by denaturation can create several million clusters of approximately 1,000 copies of single-stranded DNA molecules of the same template in each channel of the flow cell. Primers, DNA polymerase and four fluorophore-labeled, reversibly terminating nucleotides are used to perform sequential sequencing. After nucleotide incorporation, a laser is used to excite the fluorophores, and an image is captured and the identity of the first base is recorded. The 3' terminators and fluorophores from each incorporated base are removed and the incorporation, detection and identification steps are repeated. Sequencing according to this technology is described in U.S. Pat. 7,960,120; U.S. Pat. 7,835,871; U.S. Pat. 7,232,656; U.S. Pat. 7,598,035; U.S. Pat. 6,911,345; U.S. Pat. 6,833,246; U.S. Pat. 6,828,100; U.S. Pat. 6,306,597; U.S. Pat. 6,210,891;

U.S. Pub. 2011/0009278; U.S. Pub. 2007/0114362; U.S. Pub. 2006/0292611; and U.S. Pub. 2006/0024681, each of which are incorporated by reference in their entirety.

Sequencing generates a plurality of reads. Reads generally include sequences of nucleotide data wherein read length may be associated with sequencing technology. For example, the single-molecule real-time (SMRT) sequencing technology of Pacific Bio produces reads thousands of base-pairs in length. For 454 pyrosequencing, read length may be about 700 bp in length. In some embodiments, reads are less than about 500 bases in length, or less than about 150 bases in length, or less than about 90 bases in length. In certain embodiments, reads are between about 80 and about 90 bases, e.g., about 85 bases in length. In some embodiments, these are very short reads, i.e., less than about 50 or about 30 bases in length.

The sequence reads may be analyzed to characterize the target gene or region of interest. For example, mutations can be "called" (i.e., identified and reported), a haplotypte for the sample may be reported, or other analyses may be performed. Mutation calling is described in U.S. Pub. 2013/0268474. In some embodiments, an analysis may include determining copy number states of genomic regions of interest. A set of sequence reads can be analyzed by any suitable method known in the art. For example, in some embodiments, sequence reads are analyzed by hardware or software provided as part of a sequence instrument. In some embodiments, individual sequence reads are reviewed by sight (e.g., on a computer monitor). A computer program may be written that pulls an observed genotype from individual reads. In certain embodiments, analyzing the reads includes assembling the sequence reads and then genotyping the assembled reads.

Sequence assembly can be done by methods known in the art including reference-based assemblies, de novo assemblies, assembly by alignment, or combination methods. Assembly can include methods described in U.S. Pat. 8,209,130 titled Sequence Assembly by Porecca and Kennedy, the contents of each of which are hereby incorporated by reference in their entirety for all purposes. In some embodiments, sequence assembly uses the low coverage sequence assembly software (LOCAS) tool described by Klein, et al., in LOCAS-A low coverage sequence assembly tool for re-sequencing projects, PLoS One 6(8) article 23455 (2011), the contents of which are hereby incorporated by reference in their entirety. Sequence assembly is described in U.S. Pat. 8,165,821; U.S. Pat. 7,809,509; U.S. Pat. 6,223,128; U.S. Pub. 2011/0257889; and U.S. Pub. 2009/0318310, the contents of each of which are hereby incorporated by reference in their entirety.

Functions described above such as sequence read analysis or assembly can be implemented using systems of the invention that include software, hardware, firmware, hardwiring, or combinations of any of these.

FIG. 3 gives a diagram of a system 301 according to embodiments of the invention. System 301 may include an analysis instrument 303 which may be, for example, a sequencing instrument (e.g., a HiSeq 2500 or a MiSeq by Illumina). Instrument 303 includes a data acquisition module 305 to obtain results data such as sequence read data. Instrument 303 may optionally include or be operably coupled to its own, e.g., dedicated, analysis computer 333 (including an input/output mechanism, one or more processor, and memory). Additionally or alternatively, instrument 303 may be operably coupled to a server 313 or computer 349 (e.g., laptop, desktop, or tablet) via a network 309.

Computer 349 includes one or more processors and memory as well as an input/output mechanism. Where methods of the invention employ a client/server architecture, steps of methods of the invention may be performed using the server 313, which includes one or more of processors and memory, capable of obtaining data, instructions, etc., or providing results via an interface module or providing results as a file. The server 313 may be engaged over the network 309 by the computer 349 or the terminal 367, or the server 313 may be directly connected to the terminal 367, which can include one or more processors and memory, as well as an input/output mechanism.

In system 301, each computer preferably includes at least one processor coupled to a memory and at least one input/output (I/O) mechanism.

A processor will generally include a chip, such as a single core or multi-core chip, to provide a central processing unit (CPU). A process may be provided by a chip from Intel or AMD.

Memory can include one or more machine-readable devices on which is stored one or more sets of instructions (e.g., software) which, when executed by the processor(s) of any one of the disclosed computers can accomplish some or all of the methodologies or functions described herein. The software may also reside, completely or at least partially, within the main memory and/or within the processor during execution thereof by the computer system. Preferably, each computer includes a non-transitory memory such as a solid state drive, flash drive, disk drive, hard drive, etc. While the machine-readable devices can in an exemplary embodiment be a single

medium, the term "machine-readable device" should be taken to include a single medium or multiple media (e.g., a centralized or distributed database, and/or associated caches and servers) that store the one or more sets of instructions and/or data. These terms shall also be taken to include any medium or media that are capable of storing, encoding, or holding a set of instructions for execution by the machine and that cause the machine to perform any one or more of the methodologies of the present invention. These terms shall accordingly be taken to include, but not be limited to one or more solid-state memories (e.g., subscriber identity module (SIM) card, secure digital card (SD card), micro SD card, or solid-state drive (SSD)), optical and magnetic media, and/or any other tangible storage medium or media.

A computer of the invention will generally include one or more I/O device such as, for example, one or more of a video display unit (e.g., a liquid crystal display (LCD) or a cathode ray tube (CRT)), an alphanumeric input device (e.g., a keyboard), a cursor control device (e.g., a mouse), a disk drive unit, a signal generation device (e.g., a speaker), a touchscreen, an accelerometer, a microphone, a cellular radio frequency antenna, and a network interface device, which can be, for example, a network interface card (NIC), Wi-Fi card, or cellular modem.

Any of the software can be physically located at various positions, including being distributed such that portions of the functions are implemented at different physical locations.

System 301 or components of system 301 may be used to perform methods described herein. Instructions for any method step may be stored in memory and a processor may execute those instructions. System 301 or components of system 301 may be used for the analysis of genomic sequences or sequence reads (e.g., sequence assembly or variant calling).

In certain embodiments, as part of the analysis and determination of copy number states and subsequent identification of copy number variation, the sequence read counts for genomic regions of interest are normalized based on internal controls. In particular, an intra-sample normalization is performed to control for variable sequencing depths between samples. The sequence read counts for each genomic region of interest within a sample will be normalized according to the total read count across all control references within the sample.

After normalizing read counts for both the genomic regions of interest and control references, copy number states may be determined. In one embodiment, the normalized values for each sample of interest will be compared to the normalized values for a control sample. A ratio, for example, may be generated based on the comparison, wherein the ratio is indicative of

copy number and further determinative of any copy number variation. In the event that the determined copy number of a genomic region of interest of a particular sample falls within a tolerable level (as determined by ratio between test and control samples), it can be determined that genomic region of interest does not present copy number variation and thus the patient is at low risk for being a carrier of a condition or disease associated with such. In the event that the determined copy number of a genomic region of interest of a particular sample falls outside of a tolerable level, it can be determined that genomic region of interest does present copy number variation and thus the patient is at risk for being a carrier of a condition or disease associated with such.

Incorporation by Reference

References and citations to other documents, such as patents, patent applications, patent publications, journals, books, papers, web contents, have been made throughout this disclosure. All such documents are hereby incorporated herein by reference in their entirety for all purposes.

Equivalents

Various modifications of the invention and many further embodiments thereof, in addition to those shown and described herein, will become apparent to those skilled in the art from the full contents of this document, including references to the scientific and patent literature cited herein. The subject matter herein contains important information, exemplification and guidance that can be adapted to the practice of this invention in its various embodiments and equivalents thereof.

Examples

Example 1: Determination of Copy Number State of SMN1

Approximately 28 samples are collected to determine carrier status with respect to spinal muscular atrophy (SMA). Genomic DNA is extracted from whole human blood using a Gentra Puregene Blood Kit and following the Puregene protocol for DNA Purification from Whole Blood (Qiagen). Of the 28 samples, there is 1 water negative control and 7 control DNA samples

and 20 test samples. Each of the control samples includes two or more genomic regions of interest (e.g. loci) having known (or stable) copy numbers. Control samples 1-4 each include control loci and survival motor neuron genes (SMN), including telomeric SMN (SMN1) and centromeric SMN (SMN2) genes. There are a total of 17 control loci, 5 SMN1, and 5 SMN2, all of which have a known copy number of 2. Control sample 5 includes 17 control loci, each having a known copy number of 2, and 5 SMN1, each having a known copy number of 0. Control sample 6 includes 17 control loci, each having a known copy number of 2, and 5 SMN1, each having a known copy number of 1. Control sample 7 includes 17 control loci, each having a known copy number of 2, and 5 SMN1, each having a known copy number of 3 or more.

Samples are processed via method 101 to remove copies of SMN2. The sample is treated (e.g., heated) to denature genomic dsDNA. Primers specific to SMN2 that are complementary to regions flanking the SMN2 sequence are introduced. The primers are annealed to the ssDNA in the regions flanking the unwanted SMN2 segment. The annealed primers are then extended using a polymerase in a template-dependent manner to make double-stranded any single-stranded instance of SMN2 present in any sample. A double-stranded endonuclease is introduced and allowed to digest all dsDNA, thus digesting any segments that include SMN2. This stage of processing of the sample is completed by inactivating the ds endonuclease and the remaining DNA is analyzed for SMN1 by MIP capture and sequencing.

The processed samples are then fragmented and/or denatured in preparation for hybridization with molecular inversion probes. The genomic DNA of each sample is fragmented/denatured by any known method or technique sufficient to fragment genomic DNA.

Once it is isolated, MIP capture probes are hybridized to the fragmented genomic DNA in each sample by introducing capture probe mix into each sample well. In particular, the capture probe mix will generally include a plurality of SMA molecular inversion probes that are capable of binding to one or more of the genomic regions of interest (e.g., SMN1) or the control DNA. A library of molecular inversion probes is generated. The library may include a variety of different probe configurations. For example, one or more probes are capable of hybridizing specifically to the control loci and one or more probes are capable of hybridizing only to SMN1. Of those probes specific to SMN1, some are capable of producing sequences specific to that paralog while some are not capable of producing paralog-specific sequences. The library may also include one or more probes capable of hybridizing nonspecifically to both SMN1 and SMN2. However, since

SMN2 segments are removed from the sample via methods of the invention, copies of SMN2 will not interfere with analysis of SMN1.

Diluted probes are introduced to the isolated fragmented genomic DNA in each sample and the isolated whole genomic DNA is incubated in the diluted probe mix to promote hybridization. The time and temperature for incubation may be based on any known hybridization protocol, sufficient to result in hybridization of the probes to the DNA. After capture of the genomic region of interest (e.g., SMN1) the captured region is subjected to an enzymatic gap-filling and ligation step, in accordance with any known methods or techniques, including those generally described herein. The captured material may further be purified.

The purified captured DNA is then amplified by any known amplification methods or techniques. In one embodiment, the purified captured DNA is amplified using barcode-based PCR. The resulting barcodes PCRs for each sample are then combined into a master pool and quantified.

After PCR, portions of the PCR reactions for each sample are pooled and purified, then quantified. In particular, the PCR reactions for all samples are pooled in equal volumes into one master pool. The master sample pool is then purified via a PCR cleanup protocol according to manufacturer's instructions. The purified pool is then run on a microfluidics-based platform for sizing, quantification and quality control of DNA, RNA, proteins and cells. In particular, the purified pool and control samples (pre-purification) are run on an Agilent Bioanalyzer for the detection and quantification of SMN1 probe products.

Next, the sample pool is prepared for sequencing. In a preferred embodiment, Illumina sequencing techniques are used. Prior to sequencing, the sample pool is reduced to 2 nM by diluting with 1X TE. Template DNA for cluster generation is prepared by combining 10 micro-Liter of 0.1 N NaOH with 10 micro-Liter of 2 nM DNA library (sample pool) and incubating said mixture at room temperature for 5 min. The mixture is then mixed with 980 micro-Liter of HT1 buffer (Illumina), thereby reducing the denatured library to a concentration of 20 pM. This mixture is then mixed (e.g., inversion) and pulse centrifuged. Next, 225 micro-Liter of the 20 pM library is mixed with 775 micro-Liter of HT1 buffer to reduce the library pool to a concentration of 4.5 pM. The library pool having a concentration of 4.5 pM is used for on-board clustering in the sequencing.

The sequencing is carried out on the HiSeq 2500/1500 system sold by Illumina, Inc. (San Diego, CA). Sequencing is carried out with the TruSeq Rapid PE Cluster Kit and TruSeq Rapid SBS 200 cycle kit (Illumina) and in accordance with manufacturer's instructions. In addition to the reagents and mixes included within the kits, additional reagents are prepared for genomic read sequencing primers and reverse barcode sequencing primers.

The library pool undergoes sequencing under paired-end, dual-index run conditions. Sequencing generates a plurality of reads. Reads generally include sequences of nucleotide data less than about 150 bases in length, or less than about 90 bases in length. After obtaining sequence reads, they are further processed as described in U.S. Pat. 8,209,130.

Read counts for a genomic region of interest are normalized with respect to an internal control DNA. Normalized read counts are compared to the internal control DNA, thereby obtaining a ratio. A copy number state of the genomic region of interest is determined based on the comparison, specifically the ratio.

The plurality of reads generated by the sequencing method described above are analyzed to determine copy number states, and ultimately copy number variation, in any of the genomic regions of interest (e.g., SMN1) that would necessarily indicate the presence of an autosomal recessive trait in which copy number variation is diagnostic (e.g., spinal muscular atrophy). Analysis of the read counts is carried out using Illumina's HiSeq BclConverter software. Files (e.g. qSeq files) may be generated for both the genomic and barcode reads. In particular, in accordance with one method of the present invention, genomic read data for each sample is split based upon the barcode reads, which yields separate FASTQ files for each sample.

Based on the ratios, loci copy numbers may be called as follows: a ratio of <0.1 will be called a copy number state of 0; a ratio between 0.1 and 0.8 will be called a copy number state of 1; a ratio between 0.8 and 1.25 will be called a copy number state of 2; and a ratio of >1.25 will be called a copy number state of 3+.

The determined copy numbers can then be used to determine the carrier status of an individual from which the sample was obtained (i.e. whether the patient is a carrier of the disease). In particular, if the copy number state is determined to vary from the normal copy state (e.g., CN is 0, 1 or 3+), it is indicative the condition (e.g., carrier of SMA).

712

713 What is claimed is:

714

715 1. A method of removing unwanted segments of a nucleic acid from a sample, the method
716 comprising:
717 obtaining a single-stranded nucleic acid that contains an unwanted segment;
718 adding complementary nucleic acid to create a double-stranded region that contains the
719 unwanted segment; and
720 digesting the double-stranded region, thereby removing the unwanted segment from the
721 nucleic acid.

722

723 2. The method of claim 1, wherein adding the complementary nucleic acid comprises annealing
724 an oligonucleotide to the unwanted segment, thereby creating the double-stranded region.

725

726 3. The method of claim 2, wherein the annealing step comprises annealing a plurality of primers
727 to a plurality of portions of the nucleic acid that flank an unwanted segment.

728

729 4. The method of claim 1, wherein adding the complementary nucleic acid comprises:
730 annealing an oligonucleotide to a portion of the single-stranded nucleic acid that flanks
731 the unwanted segment; and
732 extending the annealed oligonucleotide to create the double-stranded region.

733

734 5. The method of claim 4, wherein the annealing step comprises annealing a plurality of primers
735 to a plurality of portions of the nucleic acid that flank an unwanted segment.

736

737 6. The method of claim 4, wherein the extending step is conducted using a polymerase enzyme
738 under conditions sufficient to cause extension of the primer in a template-dependent manner.

739

740 7. The method of claim 1, wherein the digesting step comprising exposing the sample to an
741 enzyme that preferentially digests double-stranded nucleic acid.

742

8. The method of claim 7, wherein the enzyme is selected from double-stranded endonucleases, restriction endonucleases, and nicking enzymes.

9. The method of claim 8, further comprising the step of deactivating the enzyme.

10. The method of claim 1, wherein the nucleic acid is selected from DNA, RNA, and modified nucleic acids.

11. The method of claim 1, further comprising the step of analyzing nucleic acid remaining after the digesting step.

12. The method of claim 1, wherein the digesting step results in intact genomic DNA lacking one or more unwanted segment and that is compatible with a nucleic acid analysis assay.

13. The method of claim 12, wherein the assay comprises molecular inversion probe capture.

14. The method of claim 13, wherein the assay further comprises sequencing.

15. The method of claim 14, wherein the sequencing is selected from Sanger sequencing and Next Generation Sequencing.

16. The method of claim 1, wherein the unwanted segment is a paralog, a pseudogene, or non-paralogous repetitive element.

17. The method of claim 1, further comprising the step of obtaining a sample from a subject and denaturing double-stranded DNA in the sample.

18. The method of claim 17, wherein the denaturing step comprises exposing the sample to heat, a detergent, or a basic solution.

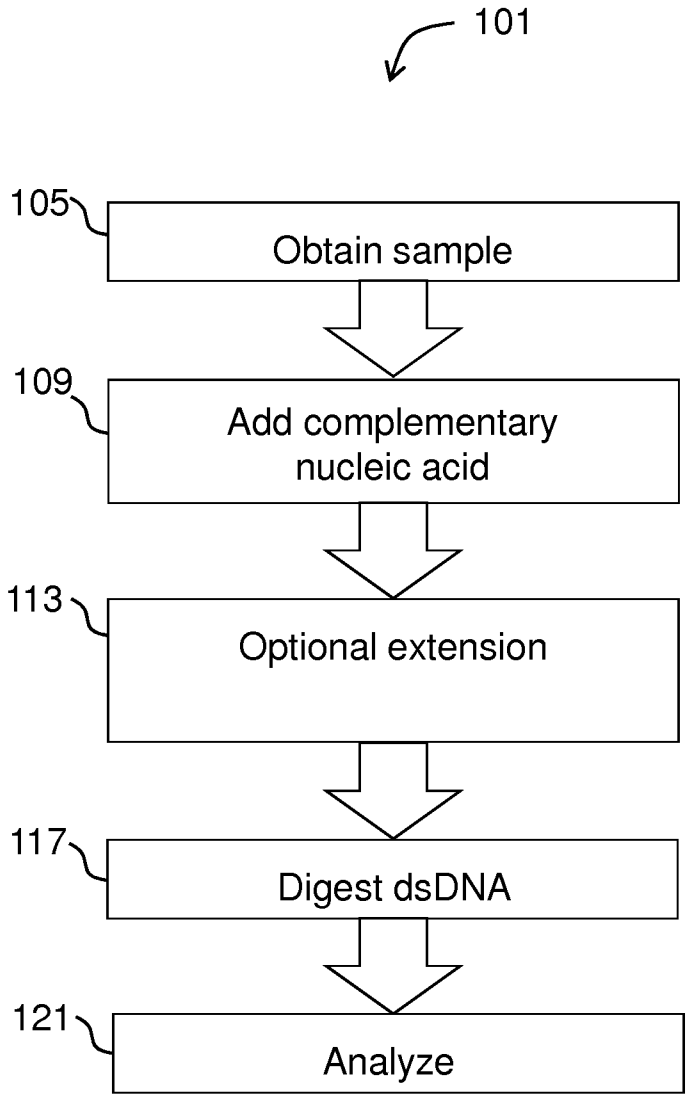
101
105
Obtain sample
109
Add complementary nucleic acid
113
Optional extension
117
Digest dsDNA
121
Analyze

FIG. 1

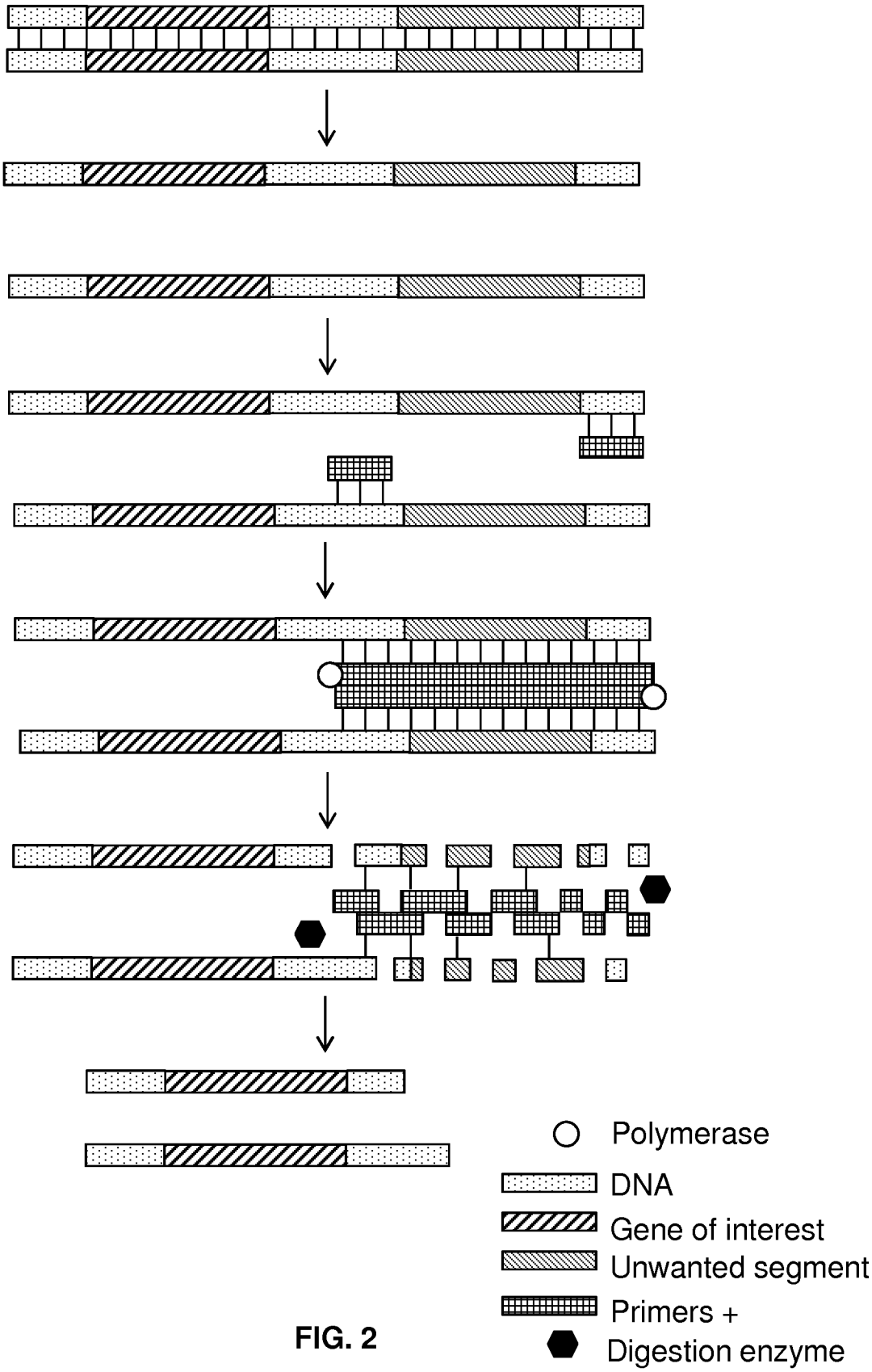
Polymerase
DNA
Gene of interest
Unwanted segment
Primers +
Digestion enzyme

FIG. 2

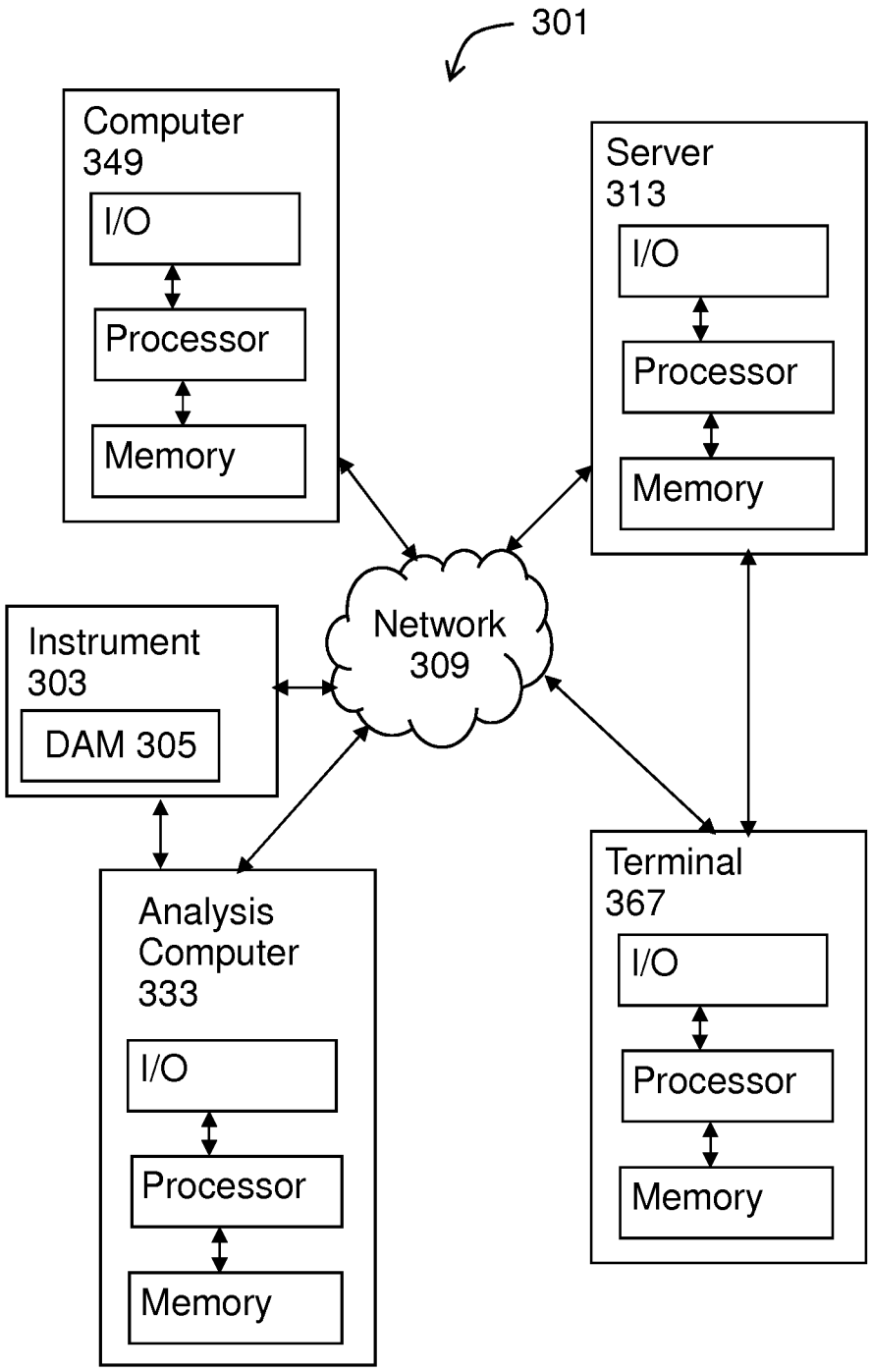
301
Computer
349
I/O
Processor
Memory
Server
313
I/O
Processor
Memory
Network
309
Instrument
303
DAM 305
Analysis
Computer
333
I/O
Processor
Memory
Terminal
367
I/O
Processor
Memory

FIG. 3

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/049132

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, BIOSIS, EMBASE, FSTA, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	S. J. GREEN ET AL: "Suicide Polymerase Endonuclease Restriction, a Novel Technique for Enhancing PCR Amplification of Minor DNA Templates", APPLIED AND ENVIRONMENTAL MICROBIOLOGY, vol. 71, no. 8, 1 August 2005 (2005-08-01) , pages 4721-4727, XP055229646, US ISSN: 0099-2240, DOI: 10.1128/AEM.71.8.4721-4727.2005	1-12, 16-18
Y	p. 4721, right-hand col., last para.; p. 4723, right-hand col., last para. - p. 4724, right-hand col., 1st para.; fig. 1, 2	13-15
	----- -/--	

[X] Further documents are listed in the continuation of Box C.

[X] 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	Date of mailing of the international search report
23 November 2015	02/12/2015

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Ripaud, Leslie

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/049132

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	STUART K ARCHER ET AL: "Selective and flexible depletion of problematic sequences from RNA-seq libraries at the cDNA stage", BMC GENOMICS, BIOMED CENTRAL LTD, LONDON, UK, vol. 15, no. 1, 26 May 2014 (2014-05-26), page 401, XP021187323, ISSN: 1471-2164, DOI: 10.1186/1471-2164-15-401	1,2, 7-11,16, 17
Y	p. 2, para. bridging left- to right-hand col.; p. 3, left-hand col., last para. - right-hand col., last para.; fig. 1	13-15

X	US 2005/003369 A1 (CHRISTIANS FREDERICK C [US] ET AL) 6 January 2005 (2005-01-06)	1,2, 7-11, 16-18
Y	para. 5-8, 47, 51-55; fig. 1A	13-15

X	WO 2013/191775 A2 (NUGEN TECHNOLOGIES INC [US]; ARMOUR CHRISTOPHER [US]; AMORESE DOUG [US) 27 December 2013 (2013-12-27) cited in the application	1-12, 16-18
Y	para. 5, 6, 12, 202; fig. 1	13-15

Y	PORRECA GREGORY J ET AL: "Multiplex amplification of large sets of human exons", NATURE METHODS, NATURE PUBLISHING GROUP, GB, vol. 4, no. 11, 1 November 2007 (2007-11-01), pages 931-936, XP002525088, ISSN: 1548-7091, DOI: 10.1038/NMETH1110 cited in the application abstract; fig. 1	13-15

X,P	WO 2015/119941 A2 (IGENOMX INTERNAT GENOMICS CORP [US]) 13 August 2015 (2015-08-13) the whole document	1-18

A	DONNA J. E. HOUSLEY ET AL: "SNP Discovery and Haplotype Analysis in the Segmentally Duplicated DRD5 Coding Region", ANNALS OF HUMAN GENETICS, vol. 73, no. 3, 1 May 2009 (2009-05-01), pages 274-282, XP055230245, GB ISSN: 0003-4800, DOI: 10.1111/j.1469-1809.2009.00513.x abstract	1-18

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2015/049132

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2005003369 A1	06-01-2005	US 2005003369 A1	06-01-2005
		US 2005123987 A1	09-06-2005
WO 2013191775 A2	27-12-2013	CA 2877094 A1	27-12-2013
		CN 104619894 A	13-05-2015
		EP 2861787 A2	22-04-2015
		GB 2518078 A	11-03-2015
		JP 2015521468 A	30-07-2015
		US 2015299767 A1	22-10-2015
		WO 2013191775 A2	27-12-2013
WO 2015119941 A2	13-08-2015	NONE	