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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
7 July 2016 (07.07.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/109364 A1

(51) International Patent Classification:
G06F 19/22 (2011.01)

(21) International Application Number:
PCT/US2015/067547

(22) International Filing Date:
28 December 2015 (28.12.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/097,139 29 December 2014 (29.12.2014) US
62/234,012 28 September 2015 (28.09.2015) US

(71) Applicant: **COUNSYL, INC.** [US/US]; 180 Kimball Way, South San Francisco, California 94080 (US).

(72) Inventors: **MUZZEY, Dale Edward**; 180 Kimball Way, South San Francisco, California 94080 (US). **ROBERTSON, Alexander De Jong**; 180 Kimball Way, South San Francisco, California 94080 (US). **EVANS, Eric Andrew**; 180 Kimball Way, South San Francisco, California 94080 (US). **MAGUIRE, Jared Robert**; 180 Kimball Way, South San Francisco, California 94080 (US).

(74) Agent: **BOYD, Victoria L.**; FisherBroyles, LLP, 555 Bryant St., #377, Palo Alto, California 94301 (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.

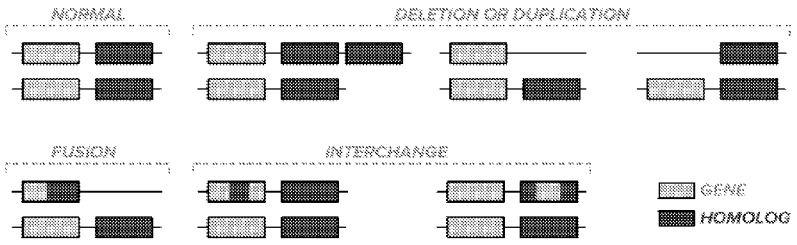
(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, 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: METHOD FOR DETERMINING GENOTYPES IN REGIONS OF HIGH HOMOLOGY

Figure 1



(57) Abstract: [1] Described herein are methods directed to determining the carrier status or genotype of a subject. Described herein is a method that combines experimental and computational approaches to resolve the structure of genomic loci (i.e., the genotype) whose sequences are highly homologous to other sequences in the genome. In particular, the determination of carrier status and/or copy number of a gene in a subject, wherein the gene has a corresponding highly homologous homolog, e.g., gene or pseudogene, utilizes Next Generation Sequencing. Also described herein is a computer-assisted method for such determinations.

WO 2016/109364 A1

METHOD FOR DETERMINING GENOTYPES IN REGIONS OF HIGH HOMOLOGY

TECHNICAL FIELD

[001] The following disclosure relates generally to determining genotypes and, more specifically, to determining genotypes associated with a gene having a corresponding highly homologous homolog.

BACKGROUND

[002] Many diseases result from genes rendered inactive by mutation. Identification of such mutations is, therefore, a fundamental goal of clinical genetic medicine. For many genes, these mutations are relatively easy to find from Next Generation Sequencing (NGS) data. However, for a subset of genes that are the subject of several important and prevalent disorders, it is challenging to identify and count the number of inactivated genes, since these genes are effectively occluded by other homologous parts of the genome.

[003] Resolving the structure and content of genomic regions that are highly homologous to other (typically dysfunctional) regions is exceptionally difficult, even with advanced NGS tools. Unfortunately, these technical obstacles are especially problematic, as many of these difficult regions have disease implications. Indeed, their very homology to dysfunctional regions leads to frequent rearrangements between genes and homologs, which can affect the number of functional copies of the gene.

[004] Thus, there remains a need for detecting and determining the genotype and/or carrier status of a subject with respect to a gene, wherein the gene has a homologous homolog.

BRIEF SUMMARY OF THE INVENTION

[005] Current technologies that allow determination of genotypes for highly homologous genes and the corresponding homologs are time- and labor-intensive, as well as expensive, making them unsuitable for widespread clinical use.

[006] The presently disclosed methods may be practiced in an affordable and high-throughput manner. Thus, there are significant time, labor and expense savings. In addition, the present method overcomes the problem of resolving structure/copy-number/genotype in regions where the unique alignment of NGS reads to genes or their homologs is compromised. Importantly, these compromising "highly homologous" regions are based on two features: (1) the length of the NGS reads in the given experiment and (2) the amount of mismatches allowed by the alignment software, *e.g.*, BWA.

[007] In an aspect there is provided herein a method for determining the genomic structure (*i.e.*, genotype) of an individual with respect to a gene of interest, wherein the gene of interest has a highly homologous homolog.

[008] In an embodiment the sequence information for the gene of interest and its homolog use primers that are directed to an exon. In certain embodiments, the sequence information is from an intron of a gene of interest and/or homolog. In certain embodiments, the sequence information is from intergenic regions.

[009] In a further embodiment, the sequence information is generated by Next Generation Sequencing (NGS). In some embodiments the NGS is high-depth whole-genome shotgun sequencing (*i.e.*, without the use of probes for enrichment). In other embodiments, the NGS is targeted sequencing such as, for example, hybrid-capture technology, multiplex amplicon enrichment, or any other means of enriching specific regions of the genome for the sequencing reaction. In some embodiments, the sequencing is done in a multiplex assay.

[0010] In one embodiment, the gene is *SMN1* and the pseudogene is *SMN2*. In an embodiment, the presence of an altered copy number of *SMN1* indicates that the subject may be a carrier for the disease spinal muscular atrophy (SMA).

[0011] In another embodiment, the gene is *CYP21A2* and the pseudogene is *CYP21A1P*. In an embodiment, the presence of an altered copy number of *CYP21A2* indicates that the subject may be a carrier for the disease congenital adrenal hyperplasia (CAH).

[0012] In an embodiment, the gene is *HBA1* and the homolog is *HBA2* (or vice versa). In an embodiment, the presence of an altered copy number of either *HBA1* or *HBA2* indicates that the subject may be a carrier for the disease alpha-thalassemia.

[0013] In a further embodiment, the gene is *GBA* and the pseudogene is *GBAP*. In an embodiment, the presence of an altered copy number of *GBA* indicates that the subject may be a carrier for the disease Gaucher's Disease.

[0014] In an embodiment, the gene is *PMS2* and the pseudogene is either *PMS2CL* or one of several other pseudogenes. As of December 2015 there were 15 pseudogenes. The pseudogenes may be selected from, but not limited to, the 13 pseudogenes known as PMS2CL with the other 12 of 13 pseudogenes numbered PMS2P1 through PMS2P12. In an embodiment, the presence of an altered copy number and/or inversions that alter orientation of the gene and pseudogene (*e.g.*, those that fuse portions of pseudogene with the gene and thus compromise gene function) may indicate that the subject has increased risk for the disease Lynch Syndrome.

[0015] In an embodiment, the gene is *CHEK2*, which has several pseudogenes. As of Dec 2014, here were seven pseudogenes. The pseudogenes may be selected from, but not limited to, *CHEK2* pseudogenes enumerated in a curated database. In an embodiment, the presence of mutations that arise from recombination with its pseudogenes—*e.g.*, a pseudogene-derived frameshift mutation—may indicate that the subject has increased risk for the disease breast cancer, among other diseases. It is well known in the art that only one

of the seven pseudogenes has been named and that risk is primarily associated with one mutation, 1100delC. However, other mutations also contribute to risk of disease. Patients are at risk for Li Fraumeni syndrome and other heritable cancers.

[0016] In an aspect, there is provided a computer system configured to execute instructions for carrying out the methods described herein.

[0017] Other objects, features and advantages of the present invention will become apparent from the following detailed description. It should be understood, however, that the detailed description and specific examples, while indicating preferred embodiments of the invention, are given by way of illustration only, since various changes and modifications within the scope and spirit of the invention will become apparent to one skilled in the art from this detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] Figure 1 illustrates various genomic structures of genes and their homologs (*e.g.*, dysfunctional homologs in the case of pseudogenes). In a "normal" sample, there are two copies each of the gene and its homolog. For many genes with homologs—indeed for the genes that underlie Gaucher's Disease, Spinal Muscular Atrophy ("SMA"), Congenital Adrenal Hyperplasia ("CAH"), and alpha-thalassemia, as well as several genes linked to various cancers—the gene and homolog are in relatively close proximity to each other on the chromosome. Some examples of chromosomes that have undergone "deletion or duplication" of the gene and/or homolog are shown. Recombination between the gene and homolog can yield "fusion" genes that are part "gene" and part "homolog". Finally, "interchange" of sequences between gene and homolog is relatively frequent.

[0019] Figure 2 is a flow chart of a method as described herein.

[0020] Figure 3 illustrates an exemplary system and environment in which various embodiments of the invention may operate.

[0021] Figure 4 illustrates an exemplary computing system.

[0022] Figure 5 is a copy number ("CN") graph of *SMN1* and *SMN2*. For 10,000 samples, we used sequencing data and the CN analysis described herein to calculate the sample's CN of *SMN1* and *SMN2* and then used these values as the x- and y-coordinates, respectively, in the scatterplot. The CN(*SMN1*), *i.e.,* the copy number of *SMN1*, of each sample was validated by an orthogonal qPCR-based assay: samples determined by this latter assay to have 1, 2, or 3 copies are indicated by circles, triangles, and squares, respectively. Note that there is very clear separation in the sequencing data between the points with CN(*SMN1*) = 1 and CN(*SMN1*) = 2. Indeed, using a cutoff of CN(*SMN1*) = 1.4 to classify samples as having either 1 or 2 copies of *SMN1*, our sequencing-based CN analysis would yield no false positives or false negatives. Other noteworthy features or the plot

include: (1) the highest density of points is near (2,2), which is the normal arrangement of the locus; (2) many samples, however, are far from (2,2), consistent with frequent conversion/deletion/duplication between *SMN1* and *SMN2*.

[0023] Figure 6 shows two copy number graphs for *GBA* and *GBAP*. For two single patient samples, CN values for *GBA* and its homolog/pseudogene *GBAP* are plotted at nine different sites, arranged from 5' to 3' (left to right). The top sample (A) is normal since it has two copies of both *GBA* and *GBAP*. However, the bottom sample (B) has undergone an "interchange" event, where the 3' end of one of the *GBAP* copies has acquired GBA-derived sequence.

[0024] Figure 7 is a copy number graph for *HBA1* and *HBA2*. The plot shows CN values for 48 patient samples in the area surrounding and including *HBA2* and *HBA1*. The thick line shows a single sample in which a large segment of a single chromosome has been deleted, hence its drop in signal for much of the right side of the figure. As expected, most of the samples have CN=2. Three samples have short deletions that occur between the Z1 and Z2 regions.

[0025] Figure 8 is a graph that shows the copy number for each probe used in the *CYP21A2* gene and its homolog *CYP21A1P*. The plots show CN values for 48 patient samples in the gene *CYP21A2* (A; left)—which affects CAH—and its pseudogene *CYP21A1P* (B; right). Each position on the x-axis is a different site in the gene, arranged from 5' to 3'. The three thick traces are samples that are known to have undergone fusion events that ablate one of the copies of the gene, hence their CN values of ~1 and ~0 in the gene plot at left. *CYP21A2* and *CYP21A1P* have undergone considerable interchange/fusion/deletion/duplication throughout evolution, which is why their traces in the plots above are more jagged than the CN traces in prior figures for Gaucher's Disease (Figure 6) and alpha-thalassemia (Figure 7). Note that one of the key goals of the CN analysis method described herein is that we want to determine the number of functional gene copies (*i.e.*, *CYP21A2* in this case). As such, we first look at sites proximal to the 5' end and use their average value to resolve CN(*CYP21A2*). Next, we consider the entirety of the trace (*i.e.*, including the 3' end) to determine what types of rearrangements have occurred.

[0026] Figure 9 is a figure illustrating how the sample data gets processed from raw read counts into values that may be interpreted for copy-number shifts. Shown are six steps and five exemplary tables (designated a, b, c, d and e) that are further described herein, *infra.*

[0027] The file of this patent contains at least one drawing in color. Copies of this patent or patent publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

DETAILED DESCRIPTION

[0028] The invention will now be described in detail by way of reference only using the following definitions and examples. All patents and publications, including all sequences disclosed within such patents and publications, referred to herein are expressly incorporated by reference.

[0029] Unless defined otherwise herein, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Singleton, *et al.,* DICTIONARY OF MICROBIOLOGY AND MOLECULAR BIOLOGY, 2D ED., John Wiley and Sons, New York (1994), and Hale & Marham, THE HARPER COLLINS DICTIONARY OF BIOLOGY, Harper Perennial, NY (1991) provide one of skill with a general dictionary of many of the terms used in this invention. Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, the preferred methods and materials are described. Practitioners are particularly directed to Sambrook *et al.,* 1989, and Ausubel FM *et al.,* 1993, for definitions and terms of the art. It is to be understood that this invention is not limited to the particular methodology, protocols, and reagents described, as these may vary.

[0030] Numeric ranges are inclusive of the numbers defining the range. The term "about" is used herein to mean plus or minus ten percent (10%) of a value. For example, "about 100" refers to any number between 90 and 110.

[0031] Unless otherwise indicated, nucleic acids are written left to right in 5' to 3' orientation; amino acid sequences are written left to right in amino to carboxy orientation, respectively.

[0032] The headings provided herein are not limitations of the various aspects or embodiments of the invention which can be had by reference to the specification as a whole. Accordingly, the terms defined immediately below are more fully defined by reference to the specification as a whole.

Definitions

[0033] As used herein, "purified" means that a molecule is present in a sample at a concentration of at least 95% by weight, or at least 98% by weight of the sample in which it is contained.

[0034] An "isolated" molecule is a nucleic acid molecule that is separated from at least one other molecule with which it is ordinarily associated, for example, in its natural environment. An isolated nucleic acid molecule includes a nucleic acid molecule contained in cells that ordinarily express the nucleic acid molecule, but the nucleic acid molecule is present extrachromasomally or at a chromosomal location that is different from its natural chromosomal location.

[0035] The term "% homology" is used interchangeably herein with the term "% identity" herein and refers to the level of nucleic acid or amino acid sequence identity between the nucleic acid sequence that encodes any one of the inventive polypeptides or the inventive polypeptide's amino acid sequence, when aligned using a sequence alignment program. In the case of a nucleic acid the term also applies to the intronic and/or intergenic regions.

[0036] For example, as used herein, 80% homology means the same thing as 80% sequence identity determined by a defined algorithm, and accordingly a homolog of a given sequence has greater than 80% sequence identity over a length of the given sequence. Exemplary levels of sequence identity include, but are not limited to, 80, 85, 90, 95, 98% or more sequence identity to a given sequence, *e.g.*, the coding sequence for any one of the inventive polypeptides, as described herein.

[0037] Exemplary computer programs which can be used to determine identity between two sequences include, but are not limited to, the suite of BLAST programs, *e.g.*, BLASTN, BLASTX, and TBLASTX, BLASTP and TBLASTN, and BLAT publicly available on the Internet. See also, Altschul, *et al*., 1990 and Altschul, *et al.,* 1997.

[0038] Sequence searches are typically carried out using the BLASTN program when evaluating a given nucleic acid sequence relative to nucleic acid sequences in the GenBank DNA Sequences and other public databases. The BLASTX program is preferred for searching nucleic acid sequences that have been translated in all reading frames against amino acid sequences in the GenBank Protein Sequences and other public databases. Both BLASTN and BLASTX are run using default parameters of an open gap penalty of 11.0, and an extended gap penalty of 1.0, and utilize the BLOSUM-62 matrix. (See, *e.g.,* Altschul, S. F., et al., Nucleic Acids Res. 25:3389-3402, 1997.)

[0039] A preferred alignment of selected sequences in order to determine "% identity" between two or more sequences, is performed using for example, the CLUSTAL-W program in MacVector version 13.0.7, operated with default parameters, including an open gap penalty of 10.0, an extended gap penalty of 0.1, and a BLOSUM 30 similarity matrix.

[0040] As used herein, “highly homologous” means that the homology between a gene and its corresponding homolog is greater than 90% over a region whose length corresponds to the NGS read length. Thus, a gene and its homolog are referred to as "highly homologous" if any region in the gene is highly homologous to the homolog. An NGS read length may range from 30nt to 400nt, from 50nt to 250nt, from 50nt to 150nt, or from 100nt to 200nt. Importantly, the entire gene's sequence need not be “highly homologous” to say a gene has a homolog; only a region in the gene needs to be highly homologous.

[0041] The term “homolog” as used herein refers to a DNA sequence that is identical or nearly identical to a gene of interest located elsewhere in the subject’s genome. The

homolog can be either another gene, a "pseudogene," or a segment of sequence that is not part of a gene.

[0042] The term “mutation” as used herein refers to both spontaneous and inherited sequence variations, including, but not limited to, variations between individuals, or between an individual’s sequence and a reference sequence. Exemplary mutations include, but are not limited to, SNPs, indel, copy number variants, inversions, translocations, chromosomal fusions, etc.

[0043] A “pseudogene” as used herein is a DNA sequence that closely resembles a gene in DNA sequence but harbors at least one change that renders it dysfunctional. The change may be a single residue mutation. The change may result in a splice variant. The change may result in early termination of translation. A pseudogene is a dysfunctional relative of a functional gene. Pseudogenes are characterized by a combination of homology to a known gene (*i.e.*, a gene of interest) and nonfunctionality.

[0044] The number of pseudogenes for genes is not limited to those enumerated herein. Pseudogenes are increasingly recognized. Therefore, a person skilled in the art would be able to determine if a sequence is a pseudogene on the basis of sequence homology or by reference to a curated database such as, for example, GeneCards (genecards.org), pseudogenes.org, *etc*.

[0045] As used herein, a “gene of interest” is a gene for which determining the number of functional copies is desired. Generally, a gene of interest has two functional copies due to the two chromosomes each having a copy of the gene of interest. The terms “gene of interest” and “gene” may be used interchangeably herein.

Process

[0046] Sequences from the region of interest are enriched, where possible, with hybrid-capture probes or PCR primers, which should be designed such that the captured and sequenced fragments contain at least one sequence that distinguishes a gene from its homolog(s). For example, hybrid-capture probes may be designed to anneal adjacent to the few bases that differ between the gene and the homolog(s)/pseudogene(s) (“diff bases”). Where such distinguishing sequence is scarce, multiple probes should be used to capture distinguishable fragments to diminish the effect of biases inherent to each particular probe’s sequence. Amplicon sequencing can be used as an alternative to hybrid-capture as a means to achieve targeted sequencing. High-depth whole-genome sequencing can be used as an alternative to targeted sequencing. Any high-throughput quantitative data that reflects the dose of a particular genomic region may be used, be it from NGS, microarrays, or any other high-throughput quantitative molecular biology technique.

[0047] The abundance of NGS sequence reads bearing gene- or homolog-derived bases permit distinction between normal (CN=2) and mutant individuals (CN≠2). Additional useful information is attainable, however, even from sequence reads that cannot distinguish gene from homolog, as in the case of *HBA1* and *HBA2*, where the normal combined CN of the two identical genes is 4, and a deletion in either gene leads to collective CN ≤ 3. Note that, in principle, the CN analysis described herein could be applied even to high-depth whole-genome shotgun sequencing (*i.e.*, without the use of probes for enrichment).

[0048] Broadly speaking, and in one example, to generate a call for a region, the following process is performed, which is illustrated as process 10 in **FIG. 2**. Initially, sequences of interest are obtained at 12. For example, reads can be collected from the bam file that overlap with the region of the call—or, critically, in the region(s) of its homolog(s)—in any way. These reads can then be clipped using their associated soft-clipping information. Auxiliary information from the aligner, *e.g.*, base-to-base alignment information, can then be discarded, and the reads become simply a sequence of bases. (In some examples, filtering based on mapping quality can be optionally performed.)

[0049] Partition reads to gene or homolog(s) based on the presence of the base(s) that distinguish them. The distinguishing base(s) exploited in this partitioning process depend on the particular gene of interest. Further, the partitioning may only use a subset of the distinguishing bases in a given read, again based on the specific application. In an embodiment where the hybrid-capture probe sequence itself becomes part of the sequenced fragment, the hybrid-capture probe is designed such that the distinguishing base is at or near the terminus of one the ends of a paired-end read. For example in such a case, the hybrid-capture probe is, *e.g.*, 39 bases long, but the sequencer reads 40 bases from the captured fragment. The probe is designed such that the 40th base is a distinguishing base, thereby allowing the entire read (*i.e.*, both ends of the paired-end read) to be partitioned to gene or homolog(s) based on the 40th position's base. The precise numbers (*i.e.*, 39 and 40) in the example above could change and yield similar results. In principle, the probe could be as short as 10bp or as long as 1000bp, though lengths in the range of 20bp-100bp are most common. In embodiments like the one above where the probe becomes part of the sequenced fragment, the sequencer must read beyond the length of the probe by at least 1bp; however, in embodiments where the captured fragment alone contains enough distinguishing bases to partition the read appropriately to gene or homolog, then sequencing need not necessarily extend beyond the length of the probe.

[0050] An exemplary treatment of experimental data is shown in **FIG. 9.** Shown is an excerpt from a table with data from a single experiment (using one Illumina flowcell). Each row is a sample. Typically, 48 or 96 samples are processed (*i.e.,* tested) in a single

experiment (*i.e.*, "Sample x" = "Sample 96"), though the analysis is valid for more or fewer samples. The analysis strongly leverages the fact that copy-number mutations are relatively rare, especially in genes associated with disease; thus, it is expected that the majority of samples will have the wild-type copy number ("CN") at each site (*i.e.,* CN = 2).

[0051] As shown in Figure 9, Table a, the sites can be partitioned into test sites (*e.g.,* "TS1", "TS2", etc.) and control sites (*e.g.,* "CS1", "CS2", etc.). The parsing of test sites (TS) versus control sites (CS) depends on the assay: for instance, in the Gaucher's disease assay, TS's are sites in *GBA* or *GBAP*, and CS's include any site in the genome for which we have data that is ***not*** in either *GBA* or *GBAP*. As another example, for the SMA test, there are only two TS sites (one for *SMN1* and the other for *SMN2*). Typically, there are several hundred CS's for each experiment. If CN analysis is done in isolation, at least 10 CS's should be used, with 50 or more being preferable (basically, you need enough sites to get a robust measurement of the median, which we'll see in Figure 9, Table b.)

[0052] The next step is depicted in Figure 9, Table b, where the median for the CS raw read values are calculated. Note that each cell in the table could contain either integer-valued raw reads or floating point number of adjusted reads, where adjustments in read number could take into account factors like sequencing bias due to GC content. Note that this is only involving the CS's, since our initial assumption is that these values have CN=2; including TS's at this point could skew the median of a given row if the row's sample has a CN mutation and TS's outnumber CS's. Unlike using the mean to represent the average, the median is robust to outlier read values which are prevalent in sequencing data; however, you still should have at least 10 CS's to get a good representation of the median. This step is effectively performed by the following equation:

$$x_{i,j} = r_{i,j} / median(r_{i,CS1} : r_{i,CSX})$$

where $r_{i,j}$ is the number of raw reads in sample *i* at site *j*. The median is evaluated over all sites *j* that are in the set of *CS* sites. $x_{i,j}$ is the "sample-normalized depth value" for sample *i* at site *j*; $x_{i,j}$ is calculated for all sites *j* in both *CS* and *TS*.

[0053] As provided for in Figure 9, the value for each cell in Table a is divided by the corresponding value for the cell's row in Table b, and the quotient is written in Table c. Now the average value across a row is ~ 1. However, further normalization is needed because there are site-specific biases in data acquisition that could corrupt our interpretation of the data. For instance, note how the values in the TSx column are systematically lower than the values in TS1 or TS2. Since it is implausible that this drop at TSx reflects a CN change in every sample (especially since the assumption is that CN variations are rare, thus it would be expected that such variations would not be in every sample), a further normalization is performed (in Figure 9, Table d) to eliminate this systematic bias.

[0054] The normalization starts with calculating the median down each column. This is done for both TS and CS columns as shown in Figure 9, Table d. Then, as shown in Figure 9, Table e, the value for each cell in Table c is divided by the corresponding value for the cell's column in Table d; then the quotient is multiplied by two, and finally the product is written in Table e. We scale the quotient by 2 since division by the average gives a normalized value centered around 1, but we know that this normalized value corresponds to a biological normal CN of 2. This step is effectively performed by the following equation:

$$CN_{i,j} = 2*x_{i,j}/median(x_{S1,j} : x_{SX,j})$$

where $x_{i,j}$ is the "sample-normalized depth value" from above. The median is calculated over all samples for site *j*. $CN_{i,j}$ is the decimal approximation of the copy number of site *j* in sample *i*. Since the copy number of a sequence in the genome is an integer value, each $CN_{i,j}$ can be rounded to its nearest integer value, and confidence in the call can be calculated as described herein.

[0055] Note that the final normalization step indicated in the equation immediately above may be modified for TS's where CN is highly variable (*i.e.*, where a small majority or even a minority of samples have CN=2). For instance, in the right plot of Figure 8, the majority of samples have CN=0—not CN=2—for TS's "WL5,B08" and "WL5,B09". We have encountered such TS's in analysis of SMA (Figure 5) and CAH (Figure 8). CN values at these challenging TS's can be determined by finding the best least-squares-deviation fit of a multimodal Gaussian distribution (with modes at empirically expected integer CN values, *e.g.*, 0, 1, 2, and 3) to the empirically observed data. The CN value for each sample can then be determined by finding the minimum distance to an integer mode of the best-fit distribution.

[0056] The final step is interpretation of the data. For each disease—Congenital Adrenal Hypertrophy (CAH), Spinal Muscular Atrophy (SMA), Gaucher's, and alpha-thalassemia—we're looking for contiguous TS's in which the CN signal deviates from 2. Note that "Sample 1" in **FIG. 9** has a CN value hovering around 1, unlike the other samples which have CN values centered at 2. These data suggest that Sample 1 has a CN mutation which has lowered its CN from two to one at the TS's. It's reassuring to see that Sample 1's CN values at CS's are ~2, suggesting that the analysis was sound (*i.e.*, it's not making the claim that the sample has a CN mutation everywhere in the genome, an implausibility).

[0057] It is worth noting that the CN analysis described herein is a critical upstream step for finding other types of clinically relevant mutations in a gene with a homolog. For instance, in addition to CN variants (shown in Figure 1), single-nucleotide polymorphisms (SNPs) may also disrupt a gene and render it dysfunctional. Standard software for recognizing SNPs uses CN as a parameter, where the expected fraction of reads bearing a SNP is 1/CN. Since most parts of the genome have CN=2, SNP-finding software by default identifies sites as

SNPs when ½ of reads contain one base (*e.g.*, C) and the other ½ have a different base (*e.g.*, T). For regions with CN variation, however, the expected fraction of reads bearing a SNP could be 1 for CN=1, 1/3 for CN=3, and so on. Critically, in the absence of a CN analysis like the one described herein, a subject who has both a SNP and CN=3 may not have the SNP identified since its representation in the data (*i.e.*, 1/3) would be less than the naively expected fraction (*i.e.*, ½). Thus, the approach we describe herein is important not only for resolving genotype in terms of CN, but also in terms of finding other mutations like SNPs and short insertions/deletions ("indels").

[0058] Since we typically have multiple TS's for a given test, we can assess confidence in our CN determination using a z-score. Here are the steps that may be used:

a. Calculate the interquartile range ("IQR") for each TS column. The IQR is the difference between the 75th- and 25th-percentile values. Assuming normal-distribution statistics, convert the IQR to a standard deviation ("SD") by dividing by ~1.33. We use the IQR as an intermediate step to finding the SD, since IQR is insensitive to outliers whereas SD can shift wildly with outliers. This attention to outliers is especially important because the rare samples with CN mutations will effectively be outliers in each column.

b. With the SD in hand for each TS column, we next enumerate the hypotheses (*i.e.*, CN=1, CN=2, etc.), and for each hypothesis we determine how many SD's away from the hypothetical CN value our observed CN values are (this number of SD's from the assumed average value is the z-score). Next, we can convert z-scores to probabilities, which allow us to assess the likelihood of the hypothesis given the data. Treating each site as an independent observation, we calculate the probability across many TS's as the product of probabilities for each TS. Our confidence score is effectively a log-odds score, where we divide the probability of the highest-probability hypothesis by the probability of the second-highest probability hypothesis, and then take the log10 of this quotient.

One of skill in the art will appreciate that other statistical approaches that are insensitive to outliers and yields an approximation of the standard deviation of the data may be used. Identification of spans of similar copy number (*e.g.*, a series of adjacent sites with CN=1, consistent with a large deletion) can be identified in a supervised manner (*e.g.*, by eye or by matching to known or hypothesized recombination sites) or unsupervised (*e.g.*, using a Hidden Markov Model).

Exemplary Architecture and Processing Environment:

[0059] An exemplary environment and system in which certain aspects and examples of the systems and processes described herein may operate. As shown in **FIG. 3**, in some examples, the system can be implemented according to a client-server model. The system can include a client-side portion executed on a user device 102 and a server-side portion executed on a server system 110. User device 102 can include any electronic device, such as a desktop computer, laptop computer, tablet computer, PDA, mobile phone (*e.g.*, smartphone), or the like.

[0060] User devices 102 can communicate with server system 110 through one or more networks 108, which can include the Internet, an intranet, or any other wired or wireless public or private network. The client-side portion of the exemplary system on user device 102 can provide client-side functionalities, such as user-facing input and output processing and communications with server system 110. Server system 110 can provide server-side functionalities for any number of clients residing on a respective user device 102. Further, server system 110 can include one or caller servers 114 that can include a client-facing I/O interface 122, one or more processing modules 118, data and model storage 120, and an I/O interface to external services 116. The client-facing I/O interface 122 can facilitate the client-facing input and output processing for caller servers 114. The one or more processing modules 118 can include various issue and candidate scoring models as described herein. In some examples, caller server 114 can communicate with external services 124, such as text databases, subscriptions services, government record services, and the like, through network(s) 108 for task completion or information acquisition. The I/O interface to external services 116 can facilitate such communications.

[0061] Server system 110 can be implemented on one or more standalone data processing devices or a distributed network of computers. In some examples, server system 110 can employ various virtual devices and/or services of third-party service providers (*e.g.*, third-party cloud service providers) to provide the underlying computing resources and/or infrastructure resources of server system 110.

[0062] Although the functionality of the caller server 114 is shown in **FIG. 3** as including both a client-side portion and a server-side portion, in some examples, certain functions described herein (*e.g.*, with respect to user interface features and graphical elements) can be implemented as a standalone application installed on a user device. In addition, the division of functionalities between the client and server portions of the system can vary in different examples. For instance, in some examples, the client executed on user device 102 can be a thin client that provides only user-facing input and output processing functions, and delegates all other functionalities of the system to a backend server.

[0063] It should be noted that server system 110 and clients 102 may further include any one of various types of computer devices, having, *e.g.*, a processing unit, a memory (which may include logic or software for carrying out some or all of the functions described herein), and a communication interface, as well as other conventional computer components (*e.g.*, input device, such as a keyboard/touch screen, and output device, such as display). Further, one or both of server system 110 and clients 102 generally includes logic (*e.g.*, http web server logic) or is programmed to format data, accessed from local or remote databases or other sources of data and content. To this end, server system 110 may utilize various web data interface techniques such as Common Gateway Interface (CGI) protocol and associated applications (or "scripts"), Java® "servlets," *i.e.*, Java® applications running on server system 110, or the like to present information and receive input from clients 102. Server system 110, although described herein in the singular, may actually comprise plural computers, devices, databases, associated backend devices, and the like, communicating (wired and/or wireless) and cooperating to perform some or all of the functions described herein. Server system 110 may further include or communicate with account servers (*e.g.*, email servers), mobile servers, media servers, and the like.

[0064] It should further be noted that although the exemplary methods and systems described herein describe use of a separate server and database systems for performing various functions, other embodiments could be implemented by storing the software or programming that operates to cause the described functions on a single device or any combination of multiple devices as a matter of design choice so long as the functionality described is performed. Similarly, the database system described can be implemented as a single database, a distributed database, a collection of distributed databases, a database with redundant online or offline backups or other redundancies, or the like, and can include a distributed database or storage network and associated processing intelligence. Although not depicted in the figures, server system 110 (and other servers and services described herein) generally include such art recognized components as are ordinarily found in server systems, including but not limited to processors, RAM, ROM, clocks, hardware drivers, associated storage, and the like (see, *e.g.*, **FIG. 4**, discussed below). Further, the described functions and logic may be included in software, hardware, firmware, or combination thereof.

[0065] **FIG. 4** depicts an exemplary computing system 600 configured to perform any one of the above-described processes, including the various calling and scoring models. In this context, computing system 600 may include, for example, a processor, memory, storage, and input/output devices (*e.g.*, monitor, keyboard, disk drive, Internet connection, *etc.*). However, computing system 600 may include circuitry or other specialized hardware for carrying out some or all aspects of the processes. In some operational settings, computing

system 600 may be configured as a system that includes one or more units, each of which is configured to carry out some aspects of the processes either in software, hardware, or some combination thereof.

[0066] FIG. 4 depicts computing system 600 with a number of components that may be used to perform the above-described processes. The main system 1402 includes a motherboard 1404 having an input/output ("I/O") section 1406, one or more central processing units ("CPU") 1408, and a memory section 1410, which may have a flash memory card 1412 related to it. The I/O section 1406 is connected to a display 1424, a keyboard 1414, a disk storage unit 1416, and a media drive unit 1418. The media drive unit 1418 can read/write a computer-readable medium 1420, which can contain programs 1422 and/or data.

[0067] At least some values based on the results of the above-described processes can be saved for subsequent use. Additionally, a non-transitory computer-readable medium can be used to store (*e.g.*, tangibly embody) one or more computer programs for performing any one of the above-described processes by means of a computer. The computer program may be written, for example, in a general-purpose programming language (*e.g.*, Pascal, C, C++, Python, Java) or some specialized application-specific language.

[0068] Various exemplary embodiments are described herein. Reference is made to these examples in a non-limiting sense. They are provided to illustrate more broadly applicable aspects of the disclosed technology. Various changes may be made and equivalents may be substituted without departing from the true spirit and scope of the various embodiments. In addition, many modifications may be made to adapt a particular situation, material, composition of matter, process, process act(s) or step(s) to the objective(s), spirit or scope of the various embodiments. Further, as will be appreciated by those with skill in the art, each of the individual variations described and illustrated herein has discrete components and features that may be readily separated from or combined with the features of any of the other several embodiments without departing from the scope or spirit of the various embodiments. All such modifications are intended to be within the scope of claims associated with this disclosure.

EXAMPLES

[0069] The present invention is described in further detain in the following examples which are not in any way intended to limit the scope of the invention as claimed. The attached Figures are meant to be considered as integral parts of the specification and description of the invention. All references cited are herein specifically incorporated by reference for all

that is described therein. The following examples are offered to illustrate, but not to limit the claimed invention.

Example 1
CALLING GENE/HOMOLOG COPY NUMBER

[0070] This example illustrates the method for determining gene/homolog copy number and is schematized in Figure 9.

[0071] The method comprises the following steps.

1. Pooled all reads that BWA (an open-source computer software program that aligns NGS reads to a reference genome) assigned to gene or homolog(s).
2. Counted depth (*i.e.*, the number of aligned reads) for gene and homolog, respectively (*e.g.*, at the intronic position that distinguishes *SMN1* from *SMN2*), based on the sequence of the read (optionally adjust read depth to take GC bias into account).
3. Tallied depth near 50 other control sites ("CS" in Figure 9)
4. Normalized each sample's gene and homolog depths by the median of the sample's 50 control depths.
5. Further adjusted the data by normalizing by each site's median value, yielding a decimal-based copy-number value (*e.g.*, 1.21).
6. Made copy-number calls (*i.e.*, mapped decimal value from prior step to an integer value) based on statistical assessment of confidence.

[0072] Results for various gene/homolog determinations are shown in Figures 5-8.

Example 2
COPY NUMBER ANALYSIS USING HYBRID-CAPTURE PROBES

[0073] This example illustrates the method for determining gene/homolog copy number for a specific gene using probes that anneal adjacent to a base that is different between the gene and the homolog(s) or pseudogene(s).

[0074] Hybrid-capture probes were designed to anneal adjacent to the few bases that differ between *CYP21A2* and *CYP21A1P* ("diff bases"). Paired-end NGS of captured fragments allows designation of reads as being either gene- or pseudogene-derived based on the diff bases. CAH variants were identified using two strategies: SNP-based calling and copy-number analysis. SNP-based calling at a given position searched for deleterious and/or pseudogene-derived bases in a pileup composed of reads with gene-derived diff bases distal from the position of interest. By contrast, copy-number analysis used read depth of diff bases to calculate the relative abundance of each variant, and deleterious variants were identified as those with excess copy number of pseudogene-derived sequence (and, conversely, depleted copy number of gene-derived sequence). Long-range PCR and Sanger sequencing were used to confirm variants in a validation study.

[0075] The test correctly identified the genotypes of positive-control samples from affected patients, and we have since run the validated CAH test on nearly 150,000 clinical samples. The variant frequencies observed are consistent with prior studies that sequenced *CYP21A2* in affected patients. There is great diversity in the copy number of gene and pseudogene: 38% of patients have at least one haplotype that does not simply have one copy of each. Evidence for recombination between gene and pseudogene is widespread, with at least 83% having a *CYP21A2* haplotype containing pseudogene-derived bases. Finally, the test identifies compound variants consistent with specific rare haplotypes, *e.g.*, (1) three copies of *CYP21A2* where one has the Q319X mutation, and (2) *CYP21A2* with a V282L mutation in cis with two copies of *CYP21A1P*, a haplotype enriched in Ashkenazi Jewish patients.

[0076] It is understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of this application and scope of the appended claims. All publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety for all purposes.

CLAIMS

What is claimed is:

1. A computer-implemented method for inferring the properties (*e.g.*, copy number, orientation, fusion-gene status, and sequence) of highly homologous genomic regions from experimental sequencing data from a genome sample relative to a reference genomic sequence, the method comprising:

    a. Obtaining NGS sequence reads experimentally from both a gene and its homolog(s) using either targeted DNA sequencing (*e.g.*, with hybrid-capture technology or amplicon sequencing using probes or primers, respectively, that are specifically designed to yield reads unique to either gene or homolog) or high-depth untargeted sequencing (*e.g.*, whole-genome shotgun sequencing);

    b. Partitioning reads *in silico* to either gene or homolog(s) based on their alignment to the human reference genome;

    c. Counting the number of reads ("depth") both at the sites of interest (*e.g.*, sites tiled across both the gene and homolog(s)), and ≥10—and preferably ≥50—control sites;

    d. Performing copy number analysis that converts raw read depth into interpretable copy-number calls via a series of normalization calculations and statistical confidence analyses; and

    e. Identifying mutations,

wherein the ability to ascertain copy number and to isolate gene-derived reads are critical parameters for proper identification of these variants.

2. The method of claim 1, wherein step (b) comprises:

    b. Partitioning reads *in silico* to either gene or homolog based on both their alignment to the human reference genome and the presence of specific base(s) that distinguish gene from homolog(s).

3. The method of claim 1, wherein step (e) comprises:

    e. Identifying mutations, which could be copy-number variants, inversions that alter orientation, gene fusions and/or short sequence variants (*e.g.*, SNPs and indels).

4. The method of claim 1, wherein the gene is *SMN1* and the pseudogene is *SMN2*.

5. The method of claim 1, wherein the gene is *CYP21A2* and the pseudogene is *CYP21A1P*.

6. The method of claim 1, wherein the gene is *HBA1* and the pseudogene is *HBA2*.

7. The method of claim 1, wherein the gene is *GBA* and the pseudogene is *GBAP*.

8. The method of claim 1, wherein the gene is *CHEK2* and the pseudogene is at least one of its pseudogenes.

9. The method of claim 1, wherein the gene is *PMS2* and the pseudogene is selected from *PMS2CL* and its other pseudogenes.

10. A non-transitory computer-readable storage medium comprising computer-executable instructions for carrying out claim 1.

11. A system comprising:

a. one or more processors;

b. memory; and

c. one or more programs, wherein the one or more programs are stored in the memory and configured to be executed by the one or more processors, the one or more programs including instructions for carrying out claim 1.

Figure 1

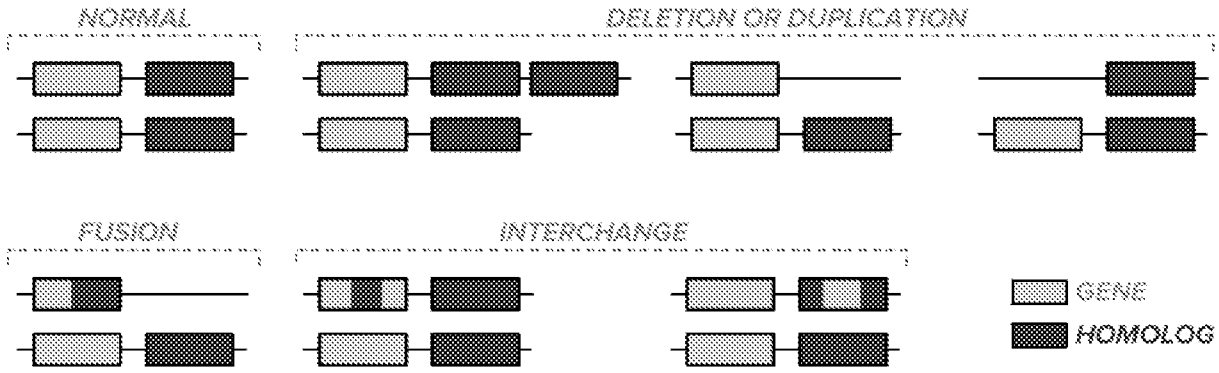
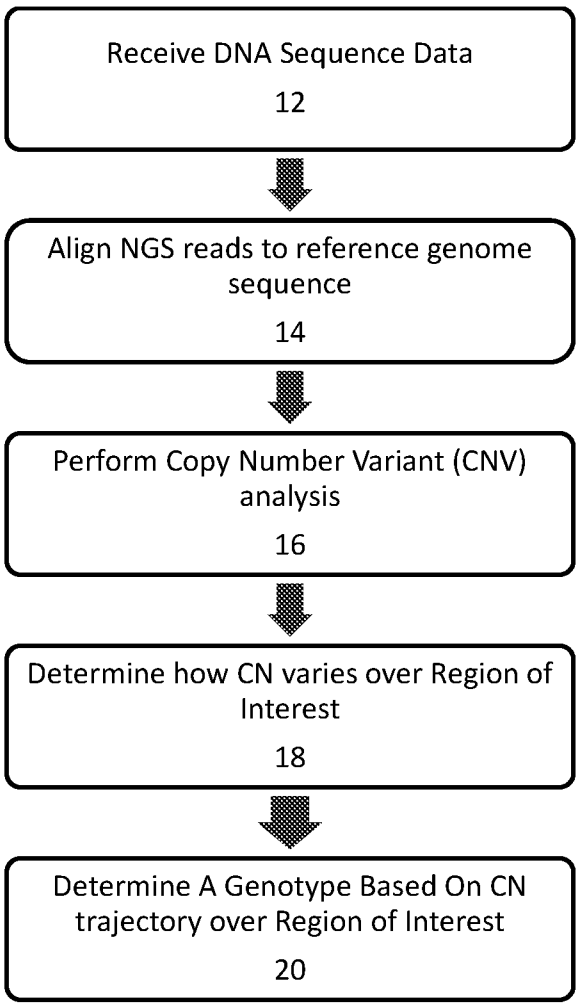

NORMAL
DELETION OR DUPLICATION
FUSION
INTERCHANGE
GENE
HOMOLOG

Figure 2

Process 10



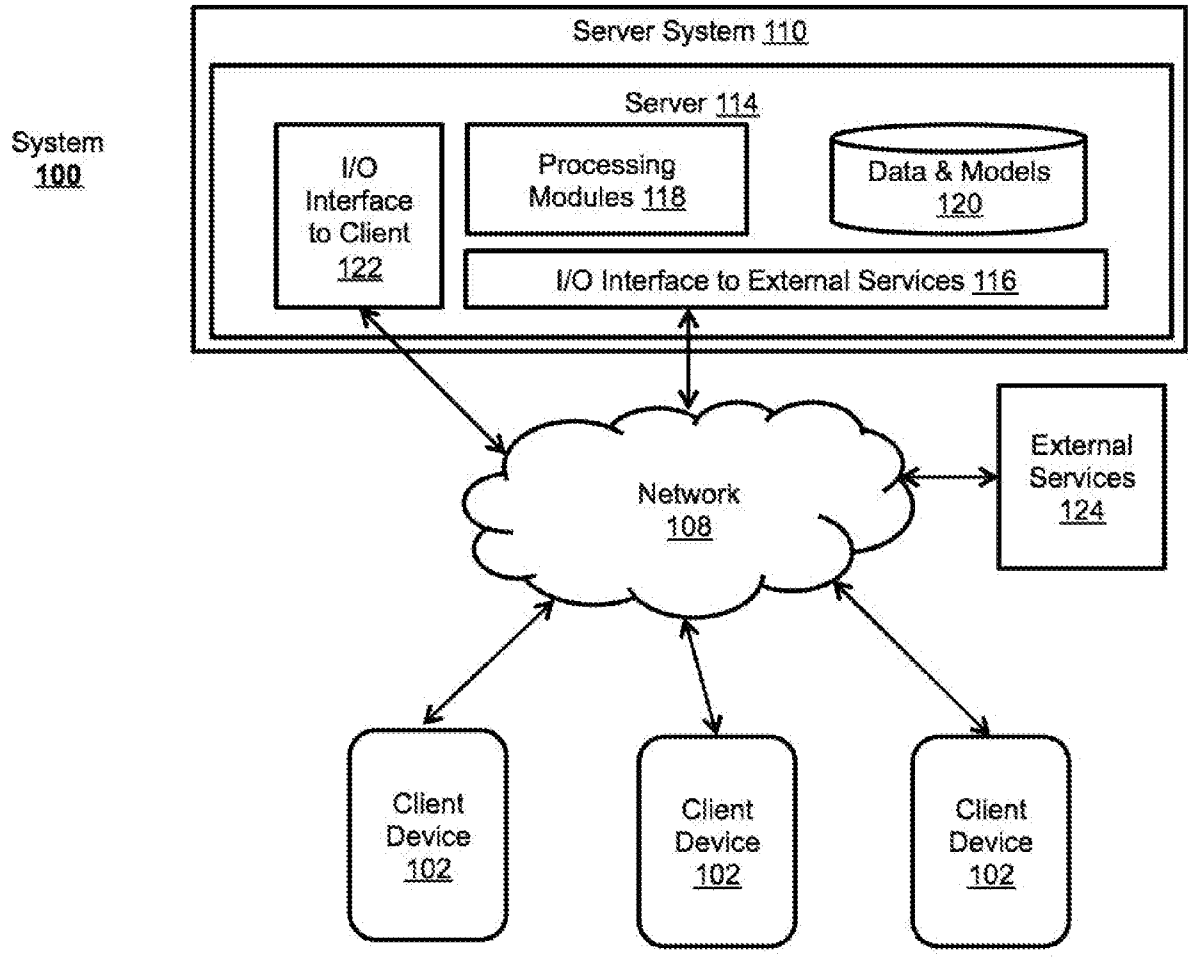

System
100
Server System 110
Server 114
I/O Interface to Client 122
Processing Modules 118
Data & Models 120
I/O Interface to External Services 116
Network 108
External Services 124
Client Device 102
Client Device 102
Client Device 102

Figure 3

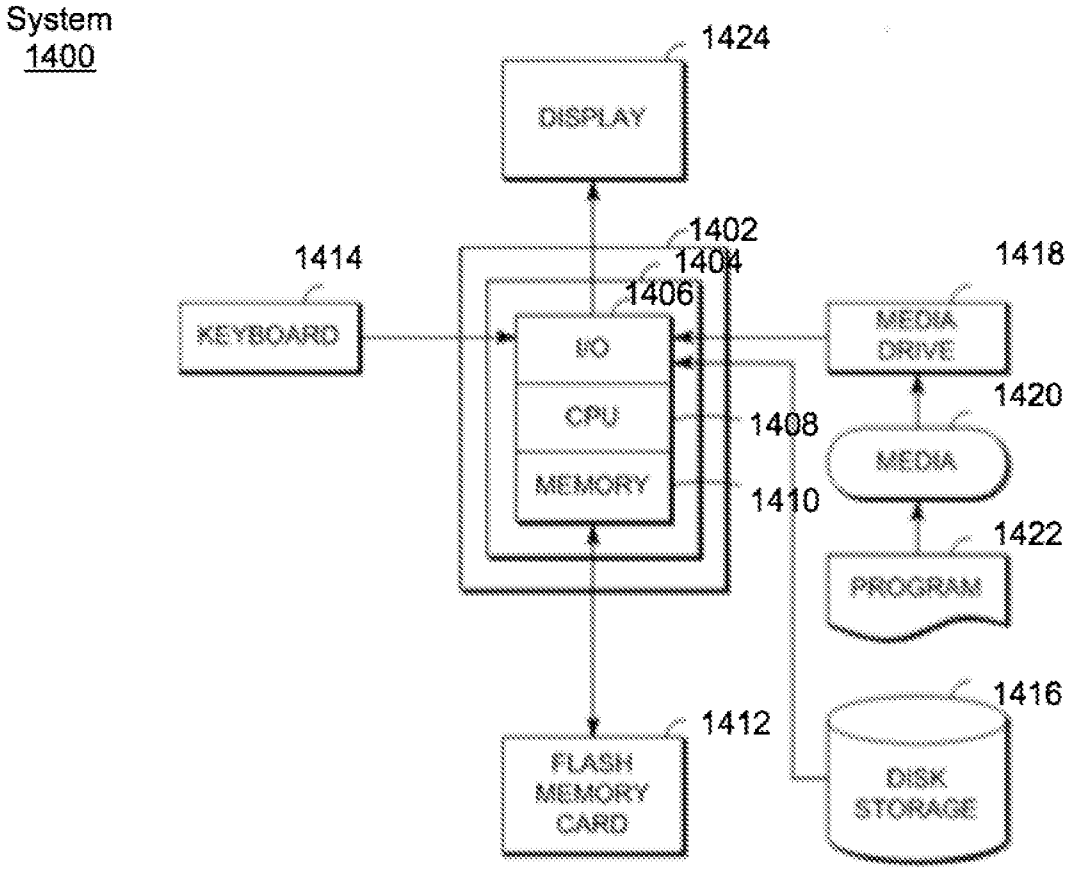

System
1400
1424
DISPLAY
1402
1404
1406
1414
KEYBOARD
I/O
CPU
MEMORY
1408
1410
1418
MEDIA DRIVE
1420
MEDIA
1422
PROGRAM
1416
DISK STORAGE
1412
FLASH MEMORY CARD

Figure 4

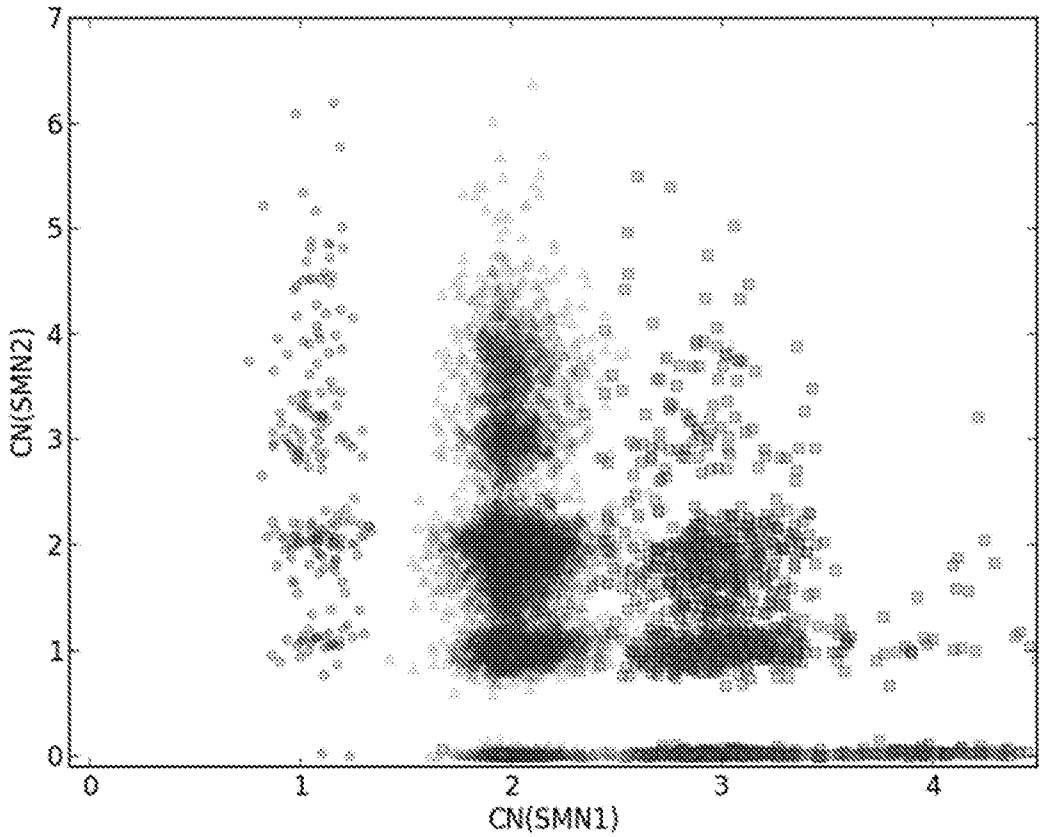
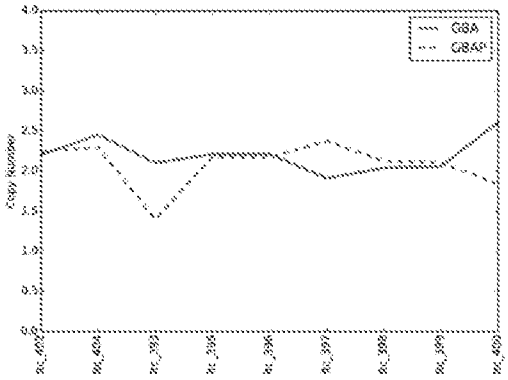


Figure 5

A



B

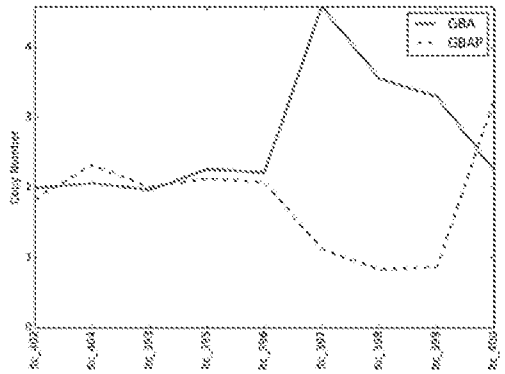


Figure 6

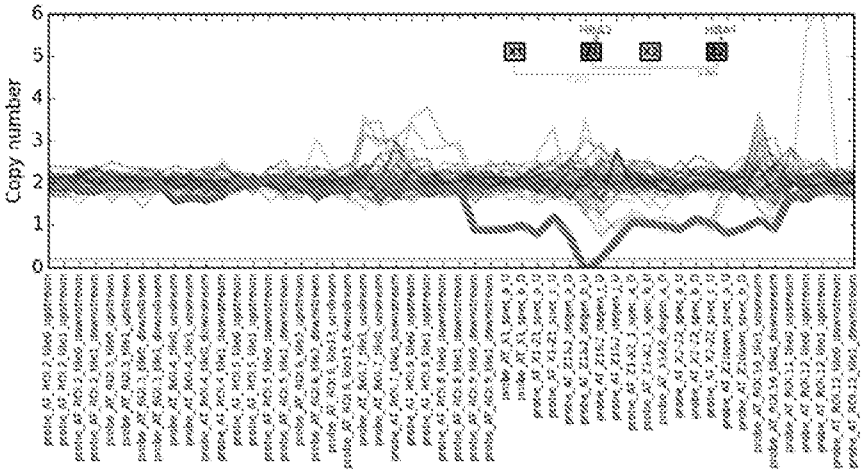


Figure 7

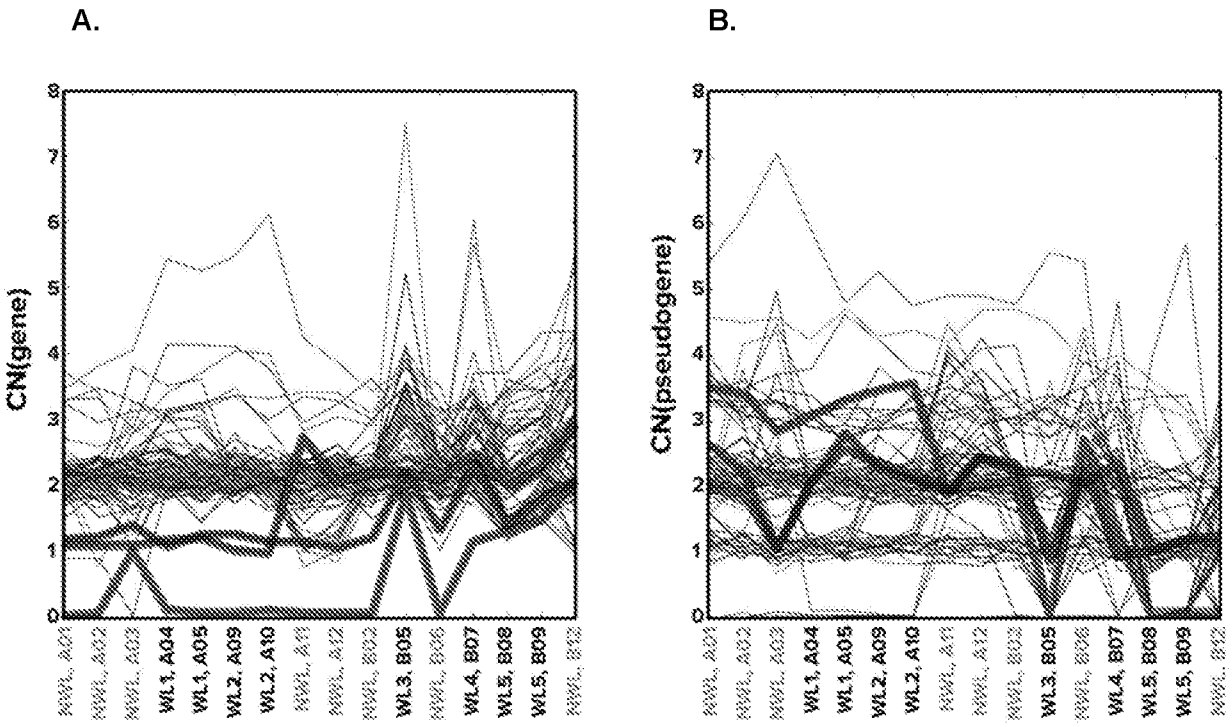

A.
B.
CN(gene)
CN(pseudogene)
0
1
2
3
4
5
6
7
8
WL1, A04
WL1, A05
WL2, A09
WL2, A10
WL3, B05
WL4, B07
WL5, B08
WL5, B09

Figure 8

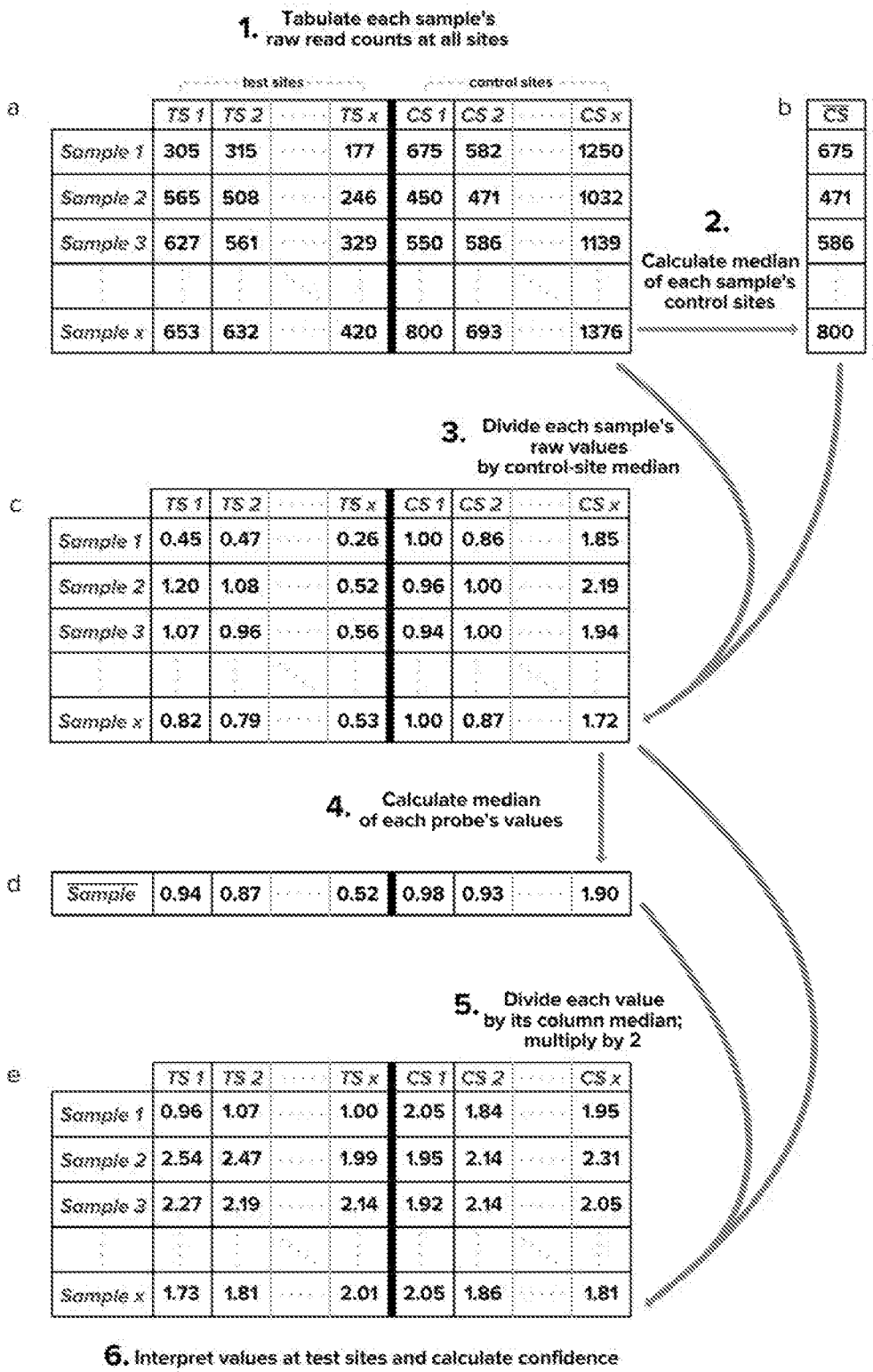

1. Tabulate each sample's raw read counts at all sites
test sites
control sites
a
TS 1 TS 2 ... TS x CS 1 CS 2 ... CS x
Sample 1 305 315 ... 177 675 582 ... 1250
Sample 2 565 508 ... 246 450 471 ... 1032
Sample 3 627 561 ... 329 550 586 ... 1139
Sample x 653 632 ... 420 800 693 ... 1376
2. Calculate median of each sample's control sites
b
CS
675
471
586
800
3. Divide each sample's raw values by control-site median
c
TS 1 TS 2 ... TS x CS 1 CS 2 ... CS x
Sample 1 0.45 0.47 ... 0.26 1.00 0.86 ... 1.85
Sample 2 1.20 1.08 ... 0.52 0.96 1.00 ... 2.19
Sample 3 1.07 0.96 ... 0.56 0.94 1.00 ... 1.94
Sample x 0.82 0.79 ... 0.53 1.00 0.87 ... 1.72
4. Calculate median of each probe's values
d
Sample 0.94 0.87 ... 0.52 0.98 0.93 ... 1.90
5. Divide each value by its column median; multiply by 2
e
TS 1 TS 2 ... TS x CS 1 CS 2 ... CS x
Sample 1 0.96 1.07 ... 1.00 2.05 1.84 ... 1.95
Sample 2 2.54 2.47 ... 1.99 1.95 2.14 ... 2.31
Sample 3 2.27 2.19 ... 2.14 1.92 2.14 ... 2.05
Sample x 1.73 1.81 ... 2.01 2.05 1.86 ... 1.81
6. Interpret values at test sites and calculate confidence

Figure 9