embodiments of the disclosure, the method may comprise: constructing a sequencing library based on a genomic DNA sample of a fetus; subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model. A formation of offspring genome equals to a random recombination with parental generation's genome (i.e., an interchange of haplotype recombination, and a random combination of gametes). For pregnant plasma, if a fetal haplotype (a recombination of parental haplotypes) is assumed as hidden states, sequencing data of the plasma may be used as observations (observing sequence), transition probabilities, observation symbol probabilities and initial state distribution may be deduced in virtue of prior data, then the most possible fetal haplotype recombination may be determined using a hidden Markov Model based on Viterbi algorithm, so as to obtain more information of fetus prior to birth. Thus, according to embodiments of the present disclosure, in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, nucleic acid sequence of a predetermined region in a fetal genome may be determined, by which a prenatal genetic detection may be effectively performed with genetic information of fetal genome.

Embodiments of a second broad aspect of the present disclosure provide a system for determining base information of a predetermined region in a fetal genome. According to embodiments of the present disclosure, the system may comprise: a library constructing apparatus, adapted for constructing sequencing library based on a genomic DNA sample of a fetus; a sequencing apparatus, connected to the library constructing apparatus, and adapted for subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and an analyzing apparatus, connected to the sequencing apparatus, and adapted for determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model. Using the system may effectively implement the above method of determining base information of a predetermined region in a fetal genome, which may determine nucleic acid sequence of a predetermined region in a fetal genome may be

Pacific Biosciences of California EXHIBIT - 1046 (Part 4) Pg. 1004
Pacific Biosciences of California v. The Chinese U. of Hong Kong - IPR2024-00028

determined in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, by which a prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

Embodiments of a third broad aspect of the present disclosure provide a computer readable medium. According to embodiments of the present disclosure, the computer readable medium including a plurality of instructions is adapted for determining base information of a predetermined region based on a sequencing result of a fetus combining with genetic information of a related individual using a hidden Markov Model. Using the computer readable medium of the present disclosure may effectively execute the plurality of instructions by a processor, to determine a nucleic acid sequence of the predetermined region in the fetal genome in virtue of the hidden Markov Model, for example using the Viterbi algorithm based on the sequencing data of the fetus combining with genetic information of a related individual, by which prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

Additional aspects and advantages of embodiments of present disclosure will be given in part in the following descriptions, become apparent in part from the following descriptions, or be learned from the practice of the embodiments of the present disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other aspects and advantages of embodiments of the present disclosure will become apparent and more readily appreciated from the following descriptions made with reference the accompanying drawings, in which:

Fig.1 is a flow chart showing an analyzing process using a hidden Markov Model according to an embodiment of the present disclosure; and

Fig.2 is a schematic diagram showing a system for determining base information of a predetermined region in a fetal genome according to an embodiment of the present disclosure.

DETAILED DESCRIPTION

Reference will be made in detail to embodiments of the present disclosure. The same or similar elements and the elements having same or similar functions are denoted

by like reference numerals throughout the descriptions. The embodiments described herein with reference to drawings are explanatory, illustrative, and used to generally understand the present disclosure. The embodiments shall not be construed to limit the present disclosure.

It should note that terms such as "first" and "second" are used herein for purposes of description and are not intended to indicate or imply relative importance or significance. Thus, features defined with “first”, “second” may explicitly or implicitly include one or more the features. Furthermore, in the description of the present disclosure, unless otherwise stated, “a/the plurality of” means two or more.

Method of determining base information of a predetermined region in a fetal genome

In a first aspect of the present disclosure, there is provided a method of determining base information of a predetermined region in a fetal genome. According to embodiments of the present disclosure, the method may comprise:

firstly, constructing a sequencing library based on a genomic DNA sample of a fetus. According to embodiments of the present disclosure, source of the genomic DNA sample of the fetus is not subjected to any special restrictions. According to some embodiments of the present disclosure, any pregnant samples containing a nucleic acid of a fetal may be used. For example, according to embodiments of the present disclosure, the pregnant sample may be breast milk, urine and peripheral blood from a pregnant woman. In which, the pregnant peripheral blood is preferred. Using the pregnant peripheral blood as the source of the genomic DNA sample of the fetus may effectively realize obtaining the genomic DNA sample of the fetus by noninvasive sampling, by which the fetal genome may be effectively monitored in the premise of having no influence on normal development of fetal growth. As for methods and processes of constructing a sequencing library for the nucleic acid sample, a person skilled in the art may appropriately select depending on different sequencing technology. Detailed process may refer to procedure provided by sequencer manufacturer, such as Illumina Company, for example, refer to Multiplexing Sample Preparation Guide (Part#1005361; Feb 2010) or Paired-End SamplePrep Guide (Part#1005063; Feb 2010) from Illumina Company, which are incorporated herein for reference. According to embodiments of the present disclosure, methods and devices for extracting a nucleic acid from a biological sample are not

subjected to any special restrictions, which may be performed using a commercial nucleic acid extracting kit.

After being constructed, obtained sequencing library is applied to a sequencer, to obtain a corresponding sequencing result consisting of a plurality of sequencing data. According to embodiments of the present disclosure, methods and devices for sequencing are not subjected to any special restrictions, including but not limited to Chain Termination Method (Sanger); a high-throughput sequencing method is preferred. Thus, using characteristics being high-throughput and deep sequencing of these apparatus, efficiency may be further improved, by which precise and accuracy of subsequent analysis with sequencing data, such as statistical test, may be further improved. The high-throughput sequencing method includes but not limited to a Next-Generation sequencing technology or a single sequencing technology. The Next-Generation sequencing platform (Metzker ML. Sequencing technologies-the next generation. *Nat Rev Genet*. 2010 Jan; 11(1):31-46) includes but not limited to Illumina-Solexa (GATM, HiSeq2000TM, etc), ABI-Solid and Roche-454 (pyrosequencing) sequencing platform; the single sequencing platform (technology) includes but not limited to True Single Molecule DNA sequencing of Helicos Company, single molecule real-time (SMRT^{TM}) of Pacific Biosciences Company, and nonapore sequencing technology of Oxford Nanopore Technologies (Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. *Nature Methods* 6 (4): 244-245), etc. With gradual development of sequencing technology, a person skilled in the art may understand other sequencing methods and apparatuses may also be used for whole genome sequencing. According to specific examples of the present disclosure, the whole genome sequencing library may be subjected to sequencing by at least one selected from Illumina-Solexa, ABI-SOLiD, Roche-454 and a single molecule sequencing apparatus.

Optionally, after being obtained, the sequencing result may be aligned to a reference sequence, to determine sequencing data corresponding to the predetermined region. Term of "predetermined region" used herein should be broadly understood, referring to any region of a nucleic acid molecule containing a possible predetermined event. For SNP analysis, it may be a region containing SNP site. For analyzing chromosome aneuploidy, the predetermined region refers to entire or part of the chromosome to be analyzed, i.e., selecting sequencing data deriving from the chromosome. Methods of

selecting sequencing data deriving from a corresponding region in the sequencing result are not subjected to any special restrictions. According to embodiments of the present disclosure, all obtained sequencing data may be aligned to a reference sequence with a known nucleic acid, to obtain the sequencing data deriving from the predetermined region. In addition, according to embodiments of the present disclosure, the predetermined region may also be a plurality of dispersal points which are not discontinuous in a genome. According to embodiments of the present disclosure, a type of used reference sequence may be not subjected to any special restrictions, which may be any known sequences contained a target region. According to embodiments of the present disclosure, the reference sequence may use a known human reference genome. For example, according to embodiments of the present disclosure, the human reference genome is NCBI 36.3, HG18. In addition, according to embodiments of the present disclosure, alignment methods are not subjected to any special restrictions. According to specific examples, SOAP may be used for alignment.

Then, determining a part of a nucleic acid sequence of the predetermined region based on sequencing data corresponding to the predetermined region; and determining other parts of the nucleic acid sequence based on determined part of the nucleic acid sequence of the predetermined region using Viterbi algorithm, to obtain the nucleic acid sequence of the predetermined region. According to embodiments of the present disclosure, the base information of the predetermined region is determined based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model. According to embodiments of the present disclosure, the base information of the predetermined region is determined using the hidden Markov Model is performed based on Viterbi algorithm. Thus, a prenatal genetic detection may be effectively performed with genetic information of fetal genome.

Referring to Fig.1, a principal for analysis using Viterbi algorithm in virtue of a hidden Markov Model is descripted in details below:

In the genetic sense, term of “a related individual” refers to individuals having a genetic relationship with a fetus. For example, according to embodiments of the present disclosure, “a related individual” may be a rental generation of a fetus, such as parents. Thus, a formation of offspring genome equals to a random recombination with parental generation’s genome (i.e., an interchange of haplotype recombination, and a random

combination of gametes). For pregnant plasma, if a fetal haplotype (a recombination of parental haplotypes) is assumed as hidden states, sequencing data of the plasma may be used as observations (observing sequence), transition probabilities, observation symbol probabilities and initial state distribution may be deduced in virtue of prior data, then the most possible fetal haplotype recombination may be determined using a hidden Markov Model based on Viterbi algorithm, so as to obtain more information of fetus prior to birth.

Steps of analyzing are show below in details:

Marker:

I. the number of sites to be detected is N.

II. haplotypes of parents are respectively recorded as $FH = \{fh_0, fh_1\}$ and $MH = \{mh_0, mh_1\}$,

in which

$mh_k = \{m_{1,k},...,m_{i,k},...,m_{N,k}\}$, $fh_k = \{f_{1,k},...,f_{i,k},...,f_{N,k}\}$,

$\forall fh_{i,k}, mh_{i,k} \in \{A,C,G,T\}$,

$k \in \{0,1\}$, $i = 1,2,3,...,N$.

III. Unknown fetal haplotype is recorded as $H = \{h_0, h_1\}$, particularly, h_0 and h_1 respectively represent inheriting from mother and father.

$h_0 = \{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$, $h_1 = \{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$,

in which $x_i \in \{0,1\}$, $y_i \in \{0,1\}$,

Subscripts x_i and y_i respectively present sequence pairs, and $q_i = \{x_i, y_i\}$ represents the hidden states which need to be decoded.

While, all hidden states possible presenting constitutes a set Q.

IV. Sequencing data is recorded as $S = \{s_1,...,s_i,...,s_N\}$

in which $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$ represents sequencing information of a site, containing the number of four bases, A, C, T and G.

V. A mean fetal concentration and a mean sequencing error rate are respectively

recorded as ε and e.

Step 1, constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination:

I. The probability distribution of the initial states is recorded as $\pi = \{\pi_j\}$ ($j \in Q$).

According to embodiments of the present disclosure, under a circumstance of having no reference data, it may assume that $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$, i.e., possibilities of each hidden state presenting at the first site are equal.

II. According to embodiments of the present disclosure, a probability of haplotype recombination is recorded as $p_r = re/N$, in which re represents a mean times of human gamete recombinations, with a prior data ranging from 25 to 30.

III. According to embodiments of the present disclosure, a transition matrix of haplotypes recombination is recorded as $A = \{a_{jk}\}$ ($j,k \in Q$), in which a_{jk} represents a probability of hidden states transition, i.e.,

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

Subscripts x_i and y_i of fetal haplotypes $h_0 = \{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$ and $h_1 = \{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$ constitute a sequence pair, $q_i = \{x_i, y_i\}$ constitute the hidden states to be encoded. For example, $x_i = 0$ represents "in a maternal chromosome, an allele in the corresponding locus is $m_{i,0}$".

Step 2, constructing a probability matrix of observations:

According to embodiments of the present disclosure, the probability matrix of observations is recorded as $B = \{b_{i,j}(s_i)\}$ ($i = 1,2,3,...,N$, $j \in Q$), in which $b_{i,j}(s_i)$ represents "an observed possibility of this sequencing information in a site i, considering maternal haplotype and fetal haplotype (state j, $j=\{x_i, y_i\}$)", i.e.,

$$b_{i,j}(s_i) = \Pr\left(s_i \middle| q_i = j, \{m_0, m_1\}\right)$$
$$= \frac{(n_{i,A} + n_{i,C} + n_{i,G} + n_{i,T})!}{n_{i,A}!n_{i,C}!n_{i,G}!n_{i,T}!} \cdot \left(P_{i,A}\right)^{n_{i,A}} \cdot \left(P_{i,C}\right)^{n_{i,C}} \cdot \left(P_{i,G}\right)^{n_{i,G}} \cdot \left(P_{i,T}\right)^{n_{i,T}},$$

in which $P_{i,base}$ represents "a possibility of a base in a site i, considering maternal haplotype and fetal haplotype (state j, $j=\{x_i, y_i\}$)", i.e.,

$$P_{i,base} = \Pr\left(base \middle| q_i = j, \{m_0, m_1\}\right)$$
$$= \sum_{k \in \{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon \cdot \Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon \cdot \Delta\left(base, f_{y_i}\right),$$

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}.$$

in which, an indicator function is

Such step is to perform HMM parameter, calculating a probability distribution of observation in each site $b_{i,j}(s_i)$, i.e., calculating a possibility presenting current sequencing data (observations) in the pregnant plasma, assuming different fetal haplotypes in each site.

Step 3, constructing a partial probability matrix, and a reversal cursor (taking an example of constructing a one-dimensional probability matrix):

Definition: partial probability $$\delta_i(q_i) = \left(\max_{q_{i-1} \in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}\right) \cdot b_{i,q_i}(s_i),$$

Definition: reversal cursor $$\Psi_i(q_i) = \arg\max_{q_{i-1} \in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}.$$

Terms of "partial probability $\delta_i(q_i)$" and "reversal cursor $\Psi_i(q_i)$" used herein both follow classic definitions of Viterbi algorithm. Detailed descriptions for the definition of the parameter may refer to Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE, Vol.77, No.2, February 1989, which is incorporated herein by reference.

Step 4, determining a final state, and tracing back an optional path

Determination of the final state, $q_N^* = \arg\max_{q_N \in Q} \delta_N(q_N)$.

The most possible fetal haplotype $q_i^* = \Psi_i(q_i)$ ($i = 1,2,3,...,N-1$) is obtained by tracing back the optional path based on the reversal curse.

Step 5, outputting a result

Thus, the sequence of the fetal genome may be effectively analyzed. Comparing to other existing method of antenatal detection, the method of the present disclosure may have following technical advantages, mainly embodying in accuracy and amount of genetic information obtainable:

1) According to embodiments of the present disclosure, a site to be detected is not limited to a parental site, for a maternal site, i.e., a maternal heterozygous site, whether a fetus inherits a maternal pathopoiesia site may also be detected excellently, with an accuracy up to 95% or more; and a plurality of abnormality types can be detected, which enlarges a range of disease detection.

2) According to embodiments of the present disclosure, information of a plurality of site and diseases may be obtained by one time of sequencing; while those gene sequence, having a low coverage in the pregnant plasma which is not able to be accurately determined only by enhancing sequencing depth, may be obtained by the method of the present disclosure, with an accurate and liable result.

3. According to embodiments of the present disclosure, a plotting with a genetic disease may be performed, some related diseases may be directly deduced with information of other sites, with a large amount of information obtained for one time, which has a more instructive meaning for clinical detection.

In addition, according to embodiments of the present disclosure, the method of

determining base information of a predetermined region in a fetal genome, not limited to a certain genetic polymorphic sites such as SNP or STR, is adapted for all genetic polymorphic sites, which may be parallel used for a plurality of sites, to verify each other. Besides applying to antenatal noninvasive detect genomic information of a fetus, achieving a purpose of disease detection, the method of the present disclosure may also be used in noninvasive antenatal paternity identification, i.e., determining an identity of a fetus' father prior birth, providing assistance for disputes involving rearing responsibilities and obligations, property and sexual assault cases, etc.

System for determining base information of a predetermined region in a fetal genome

In another aspect of the present disclosure, there is provided a system for determining base information of a predetermined region in a fetal genome. According to embodiments of the present disclosure, referring to Fig.2, the system 1000 may comprises: a library constructing apparatus 100, a sequencing apparatus 200 and an analyzing apparatus 400.

According to embodiments of the present disclosure, the library constructing apparatus 100 is adapted for constructing sequencing library based on a genomic DNA sample of a fetus. According to embodiments of the present disclosure, the sequencing apparatus 200 is connected to the library constructing apparatus 100, and adapted for subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data. According to embodiments of the present disclosure, the system 1000 may also comprise a DNA sample extracting apparatus, adapted for extracting the genomic DNA sample of the fetus from pregnant peripheral blood. Thus, the system may be adapted for noninvasive antenatal detection.

According to embodiments of the present disclosure, optionally, the system may also comprise an aligning apparatus 300. According to embodiments of the present disclosure, the aligning apparatus 300 is connected to the sequencing apparatus 200, and adapted for aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region. According to embodiments of the present disclosure, methods and devices for sequencing are not subjected to any special restrictions, including but not limited to Chain Termination

Method (Sanger); a high-throughput sequencing method is preferred. Thus, using characteristics being high-throughput and deep sequencing of these apparatus, efficiency may be further improved, by which precise and accuracy of subsequent analysis with sequencing data, such as statistical test, may be further improved. The high-throughput sequencing method includes but not limited to a Next-Generation sequencing technology or a single sequencing technology. The Next-Generation sequencing platform (Metzker ML. Sequencing technologies-the next generation. *Nat Rev Genet*. 2010 Jan; 11(1):31-46) includes but not limited to Illumina-Solexa (GATM, HiSeq2000TM, etc), ABI-Solid and Roche-454 (pyrosequencing) sequencing platform; the single sequencing platform (technology) includes but not limited to True Single Molecule DNA sequencing of Helicos Company, single molecule real-time (SMRT^{TM}) of Pacific Biosciences Company, and nonapore sequencing technology of Oxford Nanopore Technologies (Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. *Nature Methods* 6 (4): 244-245), etc. With gradual development of sequencing technology, a person skilled in the art may understand other sequencing methods and apparatuses may also be used for whole genome sequencing. According to specific examples of the present disclosure, the whole genome sequencing library may be subjected to sequencing by at least one selected from Illumina-Solexa, ABI-SOLiD, Roche-454 and a single molecule sequencing apparatus. According to embodiments of the present disclosure, a type of used reference sequence may be not subjected to any special restrictions, which may be any known sequences contained a target region. According to embodiments of the present disclosure, the reference sequence may use a known human reference genome. For example, according to embodiments of the present disclosure, the human reference genome is NCBI 36.3, HG18. In addition, according to embodiments of the present disclosure, alignment methods are not subjected to any special restrictions. According to specific examples, SOAP may be used for alignment.

According to embodiments of the present disclosure, the analyzing apparatus 400 is connected to the sequencing apparatus, and adapted for determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model.

According to embodiments of the present disclosure, in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination

probability, with re being 25~30, preferably re being 25, and N being a length of the predetermined region,

$$a_{jk} = \Pr\left(q_i = k \middle| q_{i-1} = j\right) = \begin{cases} \left(1-p_r\right)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ \left(1-p_r\right)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

is used as a recombination transition matrix with p_r being re/N.

According to embodiments of the present disclosure, the aligning apparatus is adapted for determining a base having the highest probability based on a formula of

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right)$$

wherein $\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$.

Analysis with sequencing data, which is detailed descripted above, is also adapted to the system for determining base information of a predetermined region in a fetal genome, which is omitted for brevity.

Thus, using the system may effectively implement the above method of determining base information of a predetermined region in a fetal genome, which may determine nucleic acid sequence of a predetermined region in a fetal genome may be determined in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, by which a prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

In addition, according to embodiments of the present disclosure, the predetermined region is a site previously determined as having a genetic polymorphism, and the genetic polymorphism is at least one selected from single nucleotide polymorphism and STR.

Terms of “connected” should be broadly understood, which may refer to a direct connection or indirect connection, as long as achieving the above functional connection.

It should note that a person skilled in the art may understand that features and advantages of the method of determining base information of a predetermined region in a fetal genome described above may also adapted to the system for determining base information of a predetermined region in a fetal genome, which are omitted for brevity.

Computer readable medium

In a further aspect of the present disclosure, there is provided a computer readable medium. According to embodiments of the present disclosure, the computer readable medium includes a plurality of instructions, adapted for determining base information of a predetermined region based on a sequencing result of a fetus combining with genetic information of a related individual using a hidden Markov Model. Thus, using the computer readable medium may effectively implement the above method of determining base information of a predetermined region in a fetal genome, which may determine nucleic acid sequence of a predetermined region in a fetal genome may be determined in virtue of the hidden Markov Model, for example using the Viterbi algorithm, and referring to genetic information of a related individual, by which a prenatal genetic detection may be effectively performed with genetic information of the fetal genome.

According to embodiments of the present disclosure, the plurality of instructions are adapted for determining the base information of the predetermined region using the hidden Markov model based on Viterbi algorithm. According to embodiments of the present disclosure, in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being 25~30, preferably re being 25, and N being a length of the predetermined region,

$$a_{jk} = \Pr\left(q_i = k \middle| q_{i-1} = j\right) = \begin{cases} \left(1-p_r\right)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ \left(1-p_r\right)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

is used as a recombination transition matrix with p_r being re/N.

According to embodiments of the present disclosure, the plurality of instructions are further adapted for determining a base having the highest probability based on based on a formula of

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right)$$

wherein $\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$.

Analysis with sequencing data, which is detailed descripted above, is also adapted to the computer readable medium, which is omitted for brevity.

In addition, according to embodiments of the present disclosure, the predetermined region is a site previously determined as having a genetic polymorphism, and the genetic polymorphism is at least one selected from single nucleotide polymorphism and STR.

As to the specification, "computer readable medium" may be any device adaptive for including, storing, communicating, propagating or transferring programs to be used by or in combination with the instruction execution system, device or equipment. More specific examples of the computer readable medium comprise but are not limited to: an electronic connection (an electronic device) with one or more wires, a portable computer enclosure (a magnetic device), a random access memory (RAM), a read only memory (ROM), an erasable programmable read-only memory (EPROM or a flash memory), an optical fiber device and a portable compact disk read-only memory (CDROM). In addition, the computer readable medium may even be a paper or other appropriate medium capable of printing programs thereon, this is because, for example, the paper or other appropriate medium may be optically scanned and then edited, decrypted or processed with other appropriate methods when necessary to obtain the programs in an electric manner, and then the programs may be stored in the computer memories.

It should be understood that each part of the present disclosure may be realized by the hardware, software, firmware or their combination. In the above embodiments, a plurality of steps or methods may be realized by the software or firmware stored in the memory and executed by the appropriate instruction execution system. For example, if it is realized by the hardware, likewise in another embodiment, the steps or methods may be realized by one or a combination of the following techniques known in the art: a discrete logic circuit having a logic gate circuit for realizing a logic function of a data signal, an application-specific integrated circuit having an appropriate combination logic gate circuit, a programmable gate array (PGA), a field programmable gate array (FPGA), etc.

Those skilled in the art shall understand that all or parts of the steps in the above exemplifying method of the present disclosure may be achieved by commanding the related hardware with programs. The programs may be stored in a computer readable storage medium, and the programs comprise one or a combination of the steps in the method embodiments of the present disclosure when run on a computer.

In addition, each function cell of the embodiments of the present disclosure may be

integrated in a processing module, or these cells may be separate physical existence, or two or more cells are integrated in a processing module. The integrated module may be realized in a form of hardware or in a form of software function modules. When the integrated module is realized in a form of software function module and is sold or used as a standalone product, the integrated module may be stored in a computer readable storage medium.

Reference will be made in detail to examples of the present disclosure. It would be appreciated by those skilled in the art that the following examples are explanatory, and cannot be construed to limit the scope of the present disclosure. If the specific technology or conditions are not specified in the examples, a step will be performed in accordance with the techniques or conditions described in the literature in the art (for example, referring to J. Sambrook, *et al.* (translated by Huang PT), Molecular Cloning: A Laboratory Manual, 3rd Ed., Science Press) or in accordance with the product instructions. If the manufacturers of reagents or instruments are not specified, the reagents or instruments may be commercially available, for example, from Illumina company.

General method

The method according to embodiments of the present disclosure mainly comprises following steps:

1) noninvasive sampling a pregnant sample containing fetal genetic materials, extracting genomic DNA therefrom;

2) extracting and purifying genomic DNA sample from family members of the fetus, such as parents or grandparents thereof;

3) constructing a sequencing library with every genetic material in accordance with an requirement for different sequencing platform;

4) filtering obtained sequencing data, with filtering criteria based on quality value, adaptor contamination and etc;

5) assembling obtained high-quality sequences as required, aligning an assembled result to a human genome reference sequence, to obtain uniquely-mapped sequences for analyzing using the model.

Analysis model

Marker:

I. the number of sites to be detected is N.

II. haplotypes of parents are respectively recorded as $FH = \{fh_0, fh_1\}$ and $MH = \{mh_0, mh_1\}$,

in which

$mh_k = \{m_{1,k}, ..., m_{i,k}, ..., m_{N,k}\}$, $fh_k = \{f_{1,k}, ..., f_{i,k}, ..., f_{N,k}\}$,

$\forall fh_{i,k}, mh_{i,k} \in \{A, C, G, T\}$,

$k \in \{0,1\}$, $i = 1,2,3,...,N$.

III. Unknown fetal haplotype is recorded as $H = \{h_0, h_1\}$, particularly, h_0 and h_1 respectively represent inheriting from mother and father.

$h_0 = \{m_{1,x_1}, ..., m_{i,x_i}, ..., m_{N,x_N}\}$, $h_1 = \{f_{1,y_1}, ..., f_{i,y_i}, ..., f_{N,y_N}\}$,

in which $x_i \in \{0,1\}$, $y_i \in \{0,1\}$,

Subscripts x_i and y_i respectively present sequence pairs, and $q_i = \{x_i, y_i\}$ represents the hidden states which need to be decoded.

While, all hidden states possible presenting constitutes a set Q.

IV. Sequencing data is recorded as $S = \{s_1, ..., s_i, ..., s_N\}$

in which $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$ represents sequencing information of a site, containing the number of four bases, A, C, T and G.

V. A mean fetal concentration and a mean sequencing error rate are respectively recorded as ε and e.

Step 1, constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination:

I. The probability distribution of the initial states is recorded as $\pi = \{\pi_j\}$ ($j \in Q$).

According to embodiments of the present disclosure, under a circumstance of having no reference data, it may assume that $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$, i.e., possibilities of

each hidden state presenting at the first site are equal.

II. According to embodiments of the present disclosure, a probability of haplotype recombination is recorded as $p_r = re/N$, in which re represents a mean times of human gamete recombinations, with a prior data ranging from 25 to 30.

III. According to embodiments of the present disclosure, a transition matrix of haplotypes recombination is recorded as $A = \{a_{jk}\}$ ($j,k \in Q$), in which a_{jk} represents a probability of hidden states transition, i.e.,

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

Subscripts x_i and y_i of fetal haplotypes $h_0 = \{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$ and $h_1 = \{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$ constitute a sequence pair, $q_i = \{x_i, y_i\}$ constitute the hidden states to be encoded. For example, $x_i = 0$ represents "in a maternal chromosome, an allele in the corresponding locus is $m_{i,0}$".

Step 2, constructing a probability matrix of observations:

According to embodiments of the present disclosure, the probability matrix of observations is recorded as $B = \{b_{i,j}(s_i)\}$ ($i = 1,2,3,...,N$, $j \in Q$), in which $b_{i,j}(s_i)$ represents "an observed possibility of this sequencing information in a site i, considering maternal haplotype and fetal haplotype (state j, $j=\{x_i, y_i\}$)", i.e.,

$$b_{i,j}(s_i) = \Pr(s_i | q_i = j, \{m_0, m_1\}) = \frac{(n_{i,A}+n_{i,C}+n_{i,G}+n_{i,T})!}{n_{i,A}!n_{i,C}!n_{i,G}!n_{i,T}!} \cdot (P_{i,A})^{n_{i,A}} \cdot (P_{i,C})^{n_{i,C}} \cdot (P_{i,G})^{n_{i,G}} \cdot (P_{i,T})^{n_{i,T}},$$

in which $P_{i,base}$ represents "a possibility of a base in a site i, considering maternal haplotype and fetal haplotype (state j, $j=\{x_i, y_i\}$)", i.e.,

$$P_{i,base} = \Pr(base | q_i = j, \{m_0, m_1\}) = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta(base, m_k) + \frac{1}{2}\varepsilon\cdot\Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon\cdot\Delta(base, f_{y_i}),$$

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}.$$

in which, an indicator function is

Step 3, constructing a partial probability matrix, and a reversal cursor (taking an example of constructing a one-dimensional probability matrix):

Definition: partial probability $$\delta_i(q_i)=\left(\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}\right)\cdot b_{i,q_i}(s_i)$$,

Definition: reversal cursor $$\Psi_i(q_i)=\arg\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}$$.

Step 4, determining a final state, and tracing back an optional path

Determination of the final state, $$q_N^*=\arg\max_{q_N\in Q}\delta_N(q_N)$$.

The most possible fetal haplotype $q_i^*=\Psi_i(q_i)$ ($i=1,2,3,...,N-1$) is obtained by tracing back the optional path based on the reversal curse.

Step 5, outputting a result

EXAMPLE 1

Sample collection and treatment

(1) collected sample included: peripheral blood extracted from a father and a pregnant mother within a family, and fetal umbilical cord blood after birth, all of which were collected in a tube containing EDTA for anticoagulation; saliva were collected from four grandparents using a Oragene® DNA saliva collection/DNA purification kit OG-250.

(2) extracted saliva DNA of the four grandparents were subjected to genotyping using Infinium® HD Human610-Quad BeadChip gene chip.

(3) the peripheral blood collected from the pregnant mother was centrifuged with 1600 g at 4°C for 10 min, to separate blood cells and plasma. Then obtained plasma was centrifuged with 16000 g at 4°C for 10 min, to further remove residual leukocytes, to obtain final plasma of the pregnant mother. Then genomic DNA was extracted from the final plasma of the pregnant mother using TIANamp Micro DNA Kit (TIANGEN), to obtain a genomic DNA mixture of mother and fetus thereof. Then maternal genomic DNA was extracted from removed residual leukocytes. Obtained plasma DNA were subjected to library construction based on requirement for HiSeq2000™ sequencer of Illumia® sequencer. Constructed libraries were subjected to a distribution test using Agilent®Bioanalyzer 2100 to meet a requirement for fragment ranges. Then two libraries were subjected to quantification using Q-PCR method. Qualified libraries were subjected

to sequencing using Illumina® HiSeq2000™ sequencer, with a sequencing cycle of PE101index (i.e., pair-end 101 bp index sequencing), in which parameter settings and operations were based on Illumina® specifications (obtained at http://www.illumina.com/support/documentation.ilmn)

(4) parental peripheral blood, leukocytes extracted from maternal peripheral blood and fetal umbilical cord blood were extracted with their respective genomic DNA using TIANamp Micro DNA Kit (TIANGEN).

Except for plasma DNA sample, all obtained DNA sample needed to be fragmented using Covaris™ to have a length of 500 bp. Obtained DNA fragments and plasma DNA sample were subjected to library construction based on the requirement for HiSeq2000™ sequencer of Illumia® sequencer, with a detailed procedure:

End-reparing reacting system:

10× T4 Polynucleotide kinase buffer	10 μL
dNTPs (10 mM)	4 μL
T4 DNA polymerase	5 μL
Klenow fragments	1 μL
T4 Polynucleotide kinase	5 μL
DNA fragments	30 μL
ddH_2O	up to 100 μL

After reacting at 20°C for 30 min, PCR Purification Kit (QIAGEN) was used in recycling end-repaired products. Then the recycled end-repaired products were finally dissolved in 34μL of EB buffer.

A reacting system for adding base A at end:

10× Klenow buffer	5 μL
dATP (1 mM)	10 μL
Klenow (3'-5' exo^-)	3 μL
DNA	32 μL

After incubating at 37°C for 30 min, obtained products were purified by MinElute® PCR Purification Kit (QIAGEN) and dissolved in 12 μL of EB buffer, to obtain DNA samples added with base A at end.

Ligating adaptor reacting system:

2× Rapid DNA ligating buffer	25 μL
PEI Adapter oligo-mix (20 μM)	10 μL

T4 DNA ligase	5 μL
DNA sample added with base A at end	10 μL

After reacting at 20°C for 15 min, PCR Purification Kit (QIAGEN) was used in recycling ligated products. The ligated products were finally dissolved in 32 μL of EB buffer.

PCR reacting system:

Ligated product	10 μL
Phusion DNA Polymerase Mix	25 μL
PCR primer (10 pmol/μL)	1 μL
Index N (10 pmol/μL)	1 μL
UltraPureTM Water	13 μL

Reacting procedure was shown as below:

98°C	30 s	
98°C	10 s	10 cycles
65°C	30 s	
72°C	30 s	
72°C	5 min	
4°C	Hold	

PCR Purification Kit (QIAGEN) was used in recycling PCR products, which were finally dissolved in 50 μL of EB buffer.

Constructed libraries were subjected to a distribution test using Agilent®Bioanalyzer 2100 to meet a requirement for fragment ranges. Then two libraries were subjected to quantification using Q-PCR method. Qualified libraries were subjected to sequencing using Illumina® HiSeq2000™ sequencer, with a sequencing cycle of PE101index (i.e., pair-end 101 bp index sequencing), in which parameter settings and operations were based on Illumina® specifications (obtained at http://www.illumina.com/support/documentation.ilmn)

(5) parental and maternal genomes sequencing genotyping

a. the sequencing data were aligned to a human reference genome (Hg19, NCBI 36.3) using SOAP2.

b. obtained data were subjected to consensus sequence (CNS) construction using SOAPsnp (thousands of planning data were used for Southern Han (CHS) pedigree data).

c. genotypes of a maker site were extracted.

(6) determination of parents' haplotypes

a. constructing a group genotype matrix containing ancestors' and parents' genotypes, i.e., extracting genotypes in the marker site of parents, ancestors and Southern Han pedigree.

b. deducing parents' haplotypes using BEAGLE.

(7) determination of fetal haplotype

a. aligning plasma sequencing data to a human reference genome ((Hg19, NCBI 36.3) using SOAP2;

b. constructing a probability vector of initial states, and a transition matrix of haplotypes recombination,

constructing the probability vector of initial states: taking a model of non-reference data, i.e., probabilities of every initial states were equal, being 0.25.

constructing the transition matrix of haplotypes recombination: conservatively, $re = 25$ (others were same as descriptions in "general method");

c. calculating sequencing information of each site, and constructing a probability matrix of observations (others were same as descriptions in "general method");

d. constructing a partial probability matrix, and a reversal curse (others were same as descriptions in "general method");

e. determining a final state, and tracing back an optional path; and

f. outputting.

According to genotyping results, the accuracy thereof were shown below:

			mother						total		
			homozygosis			heterozygosis					
			site number	accurate number	accuracy	site number	accurate number	accuracy	site number	accurate number	accuracy
autosome	father	homozygosis	199,552	199,552	100.00%	66,238	63,968	96.57%	265,790	263,520	99.15%
		heterozygosis	65,409	64,735	98.97%	41,849	39,944	95.45%	107,258	104,679	97.60%
			264,961	264,287	99.75%	108,087	103,912	96.14%	373,048	368,199	98.70%
chromosome X			4,881	4,881	100.00%	1,718	1,478	86.03%	6,599	6,359	96.36%

Industrial Applicability

The method of determining base information of a predetermined region in a fetal genome, the system for determining base information of a predetermined region in a fetal genome and a computer readable medium according to embodiments of the present disclosure may be effectively applied in analyzing the nucleic acid sequence of the predetermined region in the fetal genome.

Although explanatory embodiments have been shown and described, it would be appreciated by those skilled in the art that the above embodiments cannot be construed to limit the present disclosure, and changes, alternatives, and modifications can be made in the embodiments without departing from spirit, principles and scope of the present disclosure.

Reference throughout this specification to "an embodiment," "some embodiments", "one embodiment", "another example", "an example", "a specific examples", or "some examples", means that a particular feature, structure, material, or characteristic described in connection with the embodiment or example is included in at least one embodiment or example of the present disclosure. Thus, the appearances of the phrases such as "in some embodiments," "in one embodiment", "in an embodiment", "in another example, "in an example," "in a specific examples," or "in some examples," in various places throughout this specification are not necessarily referring to the same embodiment or example of the present disclosure. Furthermore, the particular features, structures, materials, or characteristics may be combined in any suitable manner in one or more embodiments or examples.

2012380221 01 Sep 2016

WHAT IS CLAIMED IS:

1. A method of determining base information of a predetermined region in a fetal genome, comprising following steps:

constructing a sequencing library based on a genomic DNA sample of a fetus;

subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

2. The method of claim 1, wherein the genomic DNA sample of the fetus is extracted from pregnant peripheral blood.

3. The method of claim 1 or claim 2, further comprising a step of aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region,

wherein optionally the reference sequence is a human reference genome,

wherein the step of aligning the sequencing result of the fetal genome to the reference sequence to determine sequencing result deriving from the predetermined region further comprises:

determining a base having the highest probability based on a formula of

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\,\Delta(base, m_k) + \frac{1}{2}\varepsilon \cdot \Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon \cdot \Delta(base, f_{y_i})$$

wherein $\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$.

4. The method of any one of claims 1 to 3, wherein the related individual is a parent of the fetus.

5. The method of any one of claims 1 to 4, wherein the step of determining the base information of the predetermined region using the hidden Markov Model is performed based on Viterbi algorithm,

wherein optionally in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being between about 25 and about 30, preferably re being 25, and N being a length of the predetermined region,

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$ is

used as a recombination transition matrix with p_r being re/N.

6. The method of any one of claims 1 to 5, wherein the predetermined region is a site previously determined as having a genetic polymorphism,

wherein optionally the genetic polymorphism is at least one selected from single nucleotide polymorphism and STR.

7. A system for determining base information of a predetermined region in a fetal genome, comprising:

a library constructing apparatus, adapted for constructing sequencing library based on a genomic DNA sample of a fetus;

a sequencing apparatus, connected to the library constructing apparatus, and

adapted for subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

an analyzing apparatus, connected to the sequencing apparatus, and adapted for determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using a hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

8. The system of claim 7, further comprising a DNA sample extracting apparatus, adapted for extracting the genomic DNA sample of the fetus from pregnant peripheral blood.

9. The system of claim 2 or claim 8, further comprising an aligning apparatus, connected to the sequencing apparatus, and adapted for aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region,

wherein the aligning apparatus is adapted for determining a base having the highest probability based on a formula of

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\,\Delta(base, m_k) + \frac{1}{2}\varepsilon\cdot\Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon\cdot\Delta(base, f_{y_i})$$

wherein

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}.$$

2012380221 01 Sep 2016

10. The system of any one of claims 7 to 9, wherein the analyzing apparatus is adapted for determining the base information of the predetermined region using a hidden Markov Model based on Viterbi algorithm, wherein optionally in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being between about 25 and about 30, preferably re being 25, and N being a length of the predetermined region,

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

is used as a recombination transition matrix with p_r being re/N.

11. A computer readable medium comprising a plurality of instructions, adapted for determining base information of a predetermined region based on a sequencing result of a fetus combining with genetic information of a related individual using a hidden Markov Model, which causes the apparatus to perform:

constructing a sequencing library based on the genomic DNA sample of the fetus;

subjecting the sequencing library to sequencing, to obtain a sequencing result of the fetus consisting of a plurality of sequencing data; and

determining the base information of the predetermined region based on the sequencing result of the fetus combining with genetic information of a related individual using the hidden Markov Model,

wherein the base information of the predetermined region comprises the most possible fetal haplotype recombination,

and the base information of the predetermined region is obtained by:

constructing a probability distribution vector of an initial state and a transition matrix of haplotypes recombination;

constructing a probability matrix of observations;

constructing a partial probability matrix, and a reversal cursor;

determining a final state, and tracing back an optional path; and

outputting a result to analyze the DNA sequence of the fetal genome.

12. The computer readable medium of claim 11, wherein the plurality of instructions are adapted for determining the base information of the predetermined region using the hidden Markov model based on Viterbi algorithm,

wherein optionally in the Viterbi algorithm, 0.25 is used as a probability distribution of an initial status, re/N is used as a recombination probability, with re being between about 25 and about 30, preferably re being 25, and N being a length of the predetermined region,

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

is used as a recombination transition matrix with p_r being re/N.

13. The computer readable medium of claim 11 or claim 12, wherein the plurality of instructions are adapted for aligning the sequencing result of the fetus to a reference sequence, to determine sequencing result deriving from the predetermined region,

wherein optionally the plurality of instructions are further adapted for determining a base having the highest probability based on based on a formula of

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\,\Delta(base, m_k) + \frac{1}{2}\varepsilon\cdot\Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon\cdot\Delta(base, f_{y_i})$$

wherein $\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$.

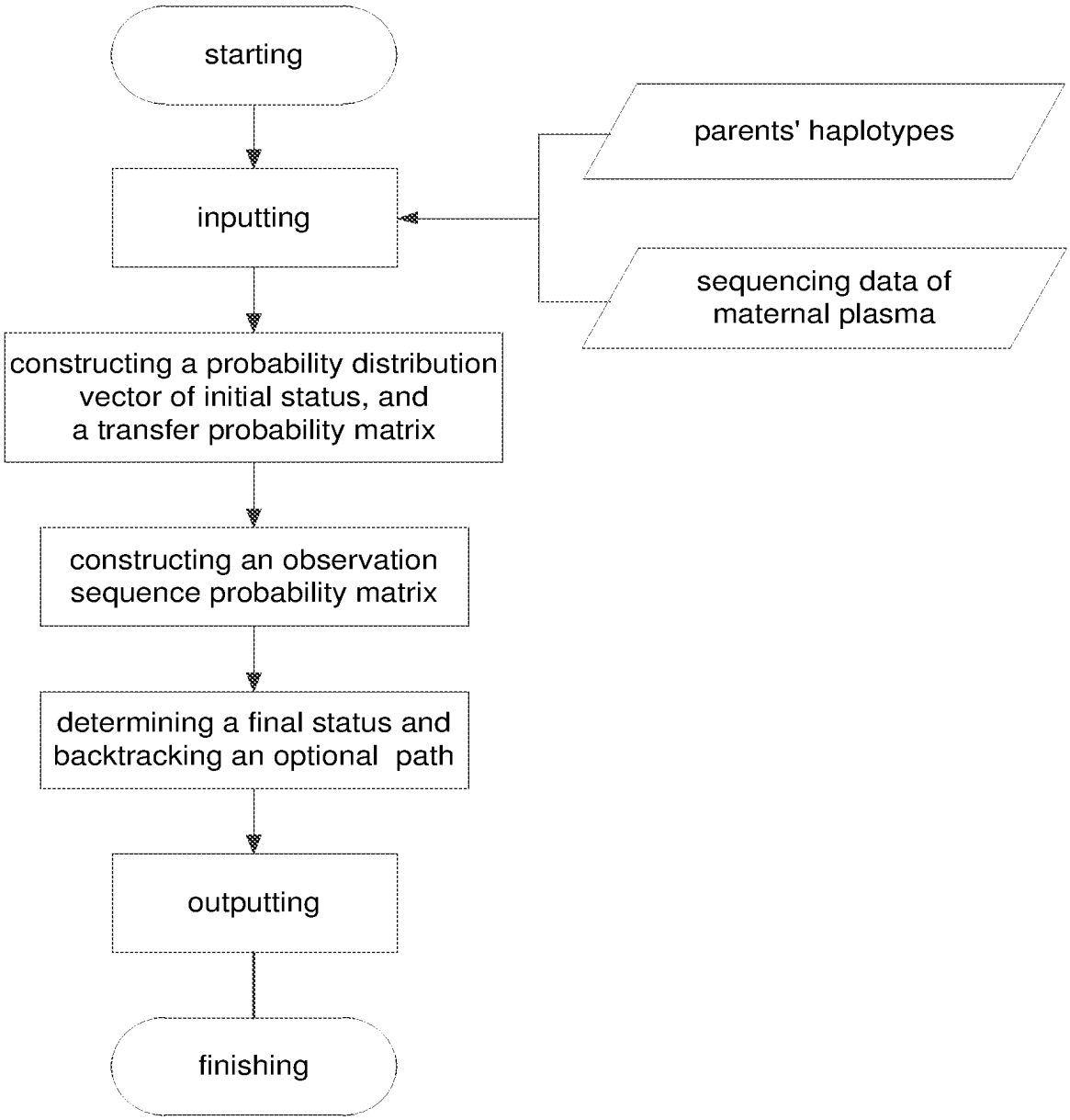

starting
inputting
parents' haplotypes
sequencing data of maternal plasma
constructing a probability distribution vector of initial status, and a transfer probability matrix
constructing an observation sequence probability matrix
determining a final status and backtracking an optional path
outputting
finishing

Fig.1

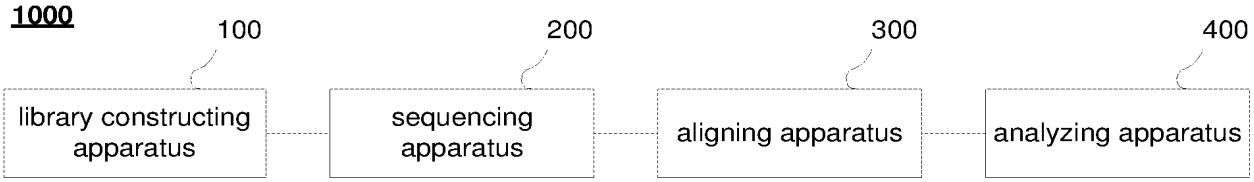

1000
100
200
300
400
library constructing apparatus
sequencing apparatus
aligning apparatus
analyzing apparatus

Fig.2

Espacenet

Bibliographic data: WO2013170429 (A1) — 2013-11-21

METHOD, SYSTEM AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION IN PREDETERMINED AREA OF FETUS GENOME

Inventor(s): CHEN SHENGPEI [CN]; GE HUIJUAN [CN]; LI XUCHAO [CN]; YI SHANG [CN]; WANG JIAN [CN]; WANG JUN [CN]; YANG HUANMING [CN]; ZHANG XIUQING [CN] ± (CHEN, SHENGPEI, ; GE, HUIJUAN, ; LI, XUCHAO, ; YI, SHANG, ; WANG, JIAN, ; WANG, JUN, ; YANG, HUANMING, ; ZHANG, XIUQING)

Applicant(s): BGI HEALTH SERVICE CO LTD [CN]; CHEN SHENGPEI [CN]; GE HUIJUAN [CN]; LI XUCHAO [CN]; YI SHANG [CN]; WANG JIAN [CN]; WANG JUN [CN]; YANG HUANMING [CN]; ZHANG XIUQING [CN] ± (BGI HEALTH SERVICE CO., LTD, ; CHEN, SHENGPEI, ; GE, HUIJUAN, ; LI, XUCHAO, ; YI, SHANG, ; WANG, JIAN, ; WANG, JUN, ; YANG, HUANMING, ; ZHANG, XIUQING)

Classification: - international: *C12Q1/68*
- cooperative: **C12Q1/6809 (CN)**; **C12Q1/6858 (EP, KR, US)**; **C12Q1/6869 (EP, KR, US)**; **C12Q1/6883 (US)**; **G06F19/18 (KR)**; **G06F19/22 (KR)**; **G06F19/24 (KR)**; **G06F19/3431 (KR)**; **G06F19/345 (KR)**; **G16B20/00 (EP, US)**; **G16B30/00 (EP, US)**; **G16H50/20 (EP, US)**; **G16H50/30 (EP, US)**; C12Q2600/118 (US); C12Q2600/156 (US) more

Application number: WO2012CN75478 20120514 Global Dossier

Priority number(s): WO2012CN75478 20120514

Also published as: AU2012380221 (A1) AU2012380221 (B2) CN104053789 (A) CN104053789 (B) EP2851431 (A1) EP2851431 (A4) EP2851431 (B1) ES2656023 (T3) HK1196401 (A1) JP2015525062 (A) JP6045686 (B2) KR101770460 (B1) KR20140146193 (A) PL2851431 (T3) RU2014150655 (A) RU2597981 (C2) SG11201407515R (A) US2015094210 (A1) US2018320235 (A1) less

Abstract of WO2013170429 (A1)

Provided are a method, system and computer readable medium for determining the base information in a predetermined area of a fetus genome, the method comprising the following steps: constructing a sequence library for the DNA samples of the fetus genome; sequencing the sequence library to obtain the sequencing result of the fetus, the sequencing result of the fetus comprised of a plurality of sequencing data; and based on the sequencing result of the fetus, determining the base information in the predetermined area according to the hidden Markov model in conjunction with the genetic information of an individual related hereditarily to the fetus.

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2013 年 11 月 21 日 (21.11.2013)

WIPO | PCT

(10) 国际公布号
WO 2013/170429 A1

(51) 国际专利分类号:
C12Q 1/68 (2006.01)

(21) 国际申请号: PCT/CN2012/075478

(22) 国际申请日: 2012 年 5 月 14 日 (14.05.2012)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人 (对除美国外的所有指定国): **深圳华大基因健康科技有限公司 (BGI HEALTH SERVICE CO., LTD.)** [CN/CN]; 中国广东省深圳市盐田区北山道 146 号北山工业区 11 栋 2、3 楼, Guangdong 518083 (CN)。

(72) 发明人; 及

(75) 发明人/申请人 (仅对美国): **陈盛培 (CHEN, Shengpei)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **葛会娟 (GE, Huijuan)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **李旭超 (LI, Xuchao)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **易赏 (YI, Shang)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **汪建 (WANG, Jian)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **王俊 (WANG, Jun)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **杨焕明 (YANG, Huanming)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

张秀清 (ZHANG, Xiuqing) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: **北京清亦华知识产权代理事务所（普通合伙） (TSINGYIHUA INTELLECTUAL PROPERTY LLC)**; 中国北京市海淀区清华园清华大学照澜院商业楼 301 室, Beijing 100084 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

(54) **Title**: METHOD, SYSTEM AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION IN PREDETERMINED AREA OF FETUS GENOME

(54) **发明名称** : 确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

(57) **Abstract**: Provided are a method, system and computer readable medium for determining the base information in a predetermined area of a fetus genome, the method comprising the following steps: constructing a sequence library for the DNA samples of the fetus genome; sequencing the sequence library to obtain the sequencing result of the fetus, the sequencing result of the fetus comprised of a plurality of sequencing data; and based on the sequencing result of the fetus, determining the base information in the predetermined area according to the hidden Markov model in conjunction with the genetic information of an individual related hereditarily to the fetus.

(57) **摘要**: 提供了确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。其中，确定胎儿基因组中预定区域碱基信息的方法，包括下列步骤：针对胎儿基因组 DNA 样本，构建测序文库；对测序文库进行测序，以便获得胎儿的测序结果，该胎儿的测序结果由多个测序数据构成；基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定预定区域的碱基信息。

WO 2013/170429 A1

确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

优先权信息

无

技术领域

本发明涉及确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。

背景技术

遗传性疾病是由于遗传物质发生改变而造成的疾病，具有先天性、家族性、终身性和遗传性的特点。遗传性疾病可分为 3 个大类：单基因遗传病、多基因遗传病及染色体异常。其中单基因病多由于单个致病基因的显性或隐性遗传所致基因功能异常；而多基因遗传病则是由多个基因变化影响所致的疾病，会在一定程度上受到外界环境因素的影响；染色体异常包括数目异常和结构异常，最为多见的是由于第 21 号染色体三体所致的唐氏综合症，患儿表现为先天愚型和肢体形状异常等其他先天性特征。由于目前对遗传性疾病尚无有效的治疗方式，只能针对性地进行支持治疗或者药物缓解，费用昂贵，给社会和家庭带来沉重经济和精神负担。因此，在孩子出生前就对孩子的患病状态进行检测，做好预防工作，以达到优生优育的目的，是十分必要的。

然而，目前的相关检测手段仍有待改进。

发明内容

本发明旨在至少解决现有技术中存在的技术问题之一。

在本发明的一个方面，本发明提出了一种确定胎儿基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：针对胎儿基因组 DNA 样本，构建测序文库；对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。对于孕期血浆，假若我们将胎儿的单倍型（父母单倍型的重组型）作为隐含状态（hidden states），可以将血浆的测序数据当做观察序列（observations），借助先验数据推算出状态转移概率（transition probabilities）、观察序列概率分布（observation symbol probabilities）和初始状态概率分布（initial state distribution），我们则可以通过诸如惠特比算法（Viterbi algorithm）根据隐马尔可夫模型推断出最可能的胎儿单倍型组合，从而获得更多胎儿的信息。因而，根据本发明的实施例，借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

在本发明的又一方面，本发明提出了一种用于确定胎儿基因组中预定区域碱基信息的系统。根据本发明的实施例，该系统包括：文库构建装置，所述文库构建装置适于针对胎儿基因组DNA样本，构建测序文库；测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；分析装置，基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。利用该系统，能够有效地实施前面所述的确定胎儿基因组中预定区域碱基信息的方法，可以借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测，从而可以有效地对胎儿基因组的遗传信息进行产前确定。

在本发明的另一方面，本发明还提出了一种计算机可读介质。根据本发明的实施例，该计算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定预定区域的碱基信息。利用本发明的计算机可读介质，能够有效地被处理器执行其存储的指令，以便借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），基于胎儿的测序结果，参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

本发明的上述和/或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

图1为根据本发明一个实施例的利用隐马尔可夫模型进行分析的流程示意图；以及

图2为根据本发明的一个实施例的用于确定胎儿基因组中预定区域核酸序列的系统的结构示意图。

发明详细描述

下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。

需要说明的是，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。进一步地，在本发明的描述中，除非另有说明，“多个”的含义是两个或两个以上。

确定胎儿基因组中预定区域碱基信息的方法

在本发明的第一方面，本发明提出了一种确定胎儿基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：

首先，针对胎儿基因组 DNA 样本，构建测序文库。根据本发明的实施例，胎儿基因组 DNA 样本的来源并不受特别限制。根据本发明的一些实施例，可以采用任何含有胎儿核酸的孕妇样本。例如，根据本发明的实施例，可以采用的孕妇样本为孕妇乳汁、尿液和外周血。其中，优选孕妇外周血。采用孕妇外周血作为胎儿基因组 DNA 样本的来源，可以有效地实现无创取样方式获得胎儿基因组 DNA，从而可以在不影响胎儿正常发育的前提下，对胎儿的基因组进行有效监测。关于针对核酸样本，构建测序文库的方法和流程，本领域技术人员可以根据不同的测序技术进行适当选择，关于流程的细节，可以参见测序仪器的厂商例如 Illumina 公司所提供的规程，例如参见 Illumina 公司 Multiplexing Sample Preparation Guide（Part#1005361; Feb 2010）或 Paired-End SamplePrep Guide（Part#1005063; Feb 2010），通过参照将其并入本文。根据本发明的实施例，从生物样本提取核酸样本的方法和设备，也不受特别限制，可以采用商品化的核酸提取试剂盒进行。

在构建测序文库后，将测序文库应用于测序仪器，对测序文库进行测序，并获得相应的测序结果，该测序结果是由多个测序数据构成的。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，能够提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46）包括但不限于 Illumina-Solexa（GA™,HiSeq2000™ 等）、ABI-Solid 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing）， Pacific Biosciences 公司单分子实时测序（single molecule real-time (SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245）。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。

任选地，在得到测序结果之后，可以将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据。在本文中所使用的术语“预定区域”应作广义理解，是指任何包含可能发生预定事件位点的核酸分子的区域。对于 SNP 分析而言，可以是指包含 SNP 位点的区域。对于分析染色体非整倍性，则预定区域指的是所要分析的染色体的全长或者部分，即选择所有来自该染色体的测序数据。从测序结果中选择来自相应区域的测序数据的方法可以不受特别限制。根据本发明的实施例，可以通过将所得到的所有测序数据与已知的核酸参照序列进行比对，从而得到来自于预定区域的测序数据。另外，根据本发

明的实施例，预定区域也可以是基因组上不连续的多个分散点。根据本发明的实施例，可以使用的参照序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参照序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI 36.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

接下来，基于与预定区域对应的测序数据，确定预定区域中的部分核酸序列；以及基于所确定的预定区域中的部分核酸序列，按照惠特比算法，确定预定区域的其他核酸序列，以便获得预定区域的核酸序列。根据本发明的实施例，可以通过基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，可以利用惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息。由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

下面参考图 1，对利用惠特比算法借助隐马尔可夫模型进行分析的原理进行详细描述：

在本文中所使用的术语“胎儿遗传相关个体”指的是在遗传意义上，与胎儿之间具有亲缘关系的个体，例如根据本发明的实施例，可以采用的“胎儿遗传相关个体”为胎儿的亲代例如父母。由此，子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。对于孕期血浆，假若我们将胎儿的单倍型（父母单倍型的重组型）作为隐含状态（hidden states），可以将血浆的测序数据当做观察序列（observations），借助先验数据推算出状态转移概率（transition probabilities）、观察序列概率分布（observation symbol probabilities）和初始状态概率分布（initial state distribution），我们则可以通过惠特比算法（Viterbi algorithm）推断出最可能的胎儿单倍型组合，从而获得更多胎儿的信息。

详细分析步骤如下：

记号：

I.需要检测的位点数为 N。

II. 父母的单倍型分别记为 $FH=\{fh_0,fh_1\}$ 和 $MH=\{mh_0,mh_1\}$，

其中，

$mh_k=\{m_{1,k},...,m_{i,k},...,m_{N,k}\}$，$fh_k=\{f_{1,k},...,f_{i,k},...,f_{N,k}\}$，

$\forall fh_{i,k},mh_{i,k}\in\{A,C,G,T\}$，

$k\in\{0,1\}$，$i=1,2,3,...,N$。

III. 将未知的胎儿单倍型记为 $H=\{h_0,h_1\}$，特别地，h_0 和 h_1 分别遗传自母亲和父亲。

$h_0=\{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$，$h_1=\{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$，

其中， $x_i \in \{0,1\}$， $y_i \in \{0,1\}$，

下标 x_i 和 y_i 组成的序列对， $q_i = \{x_i, y_i\}$ 组成了我们需要解码的隐藏状态，

而所有可能出现的隐藏状态组成了集合 Q。

IV. 测序数据记为， $S = \{s_1, ..., s_i, ..., s_N\}$，

其中， $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$，代表此位点的测序信息，包含了 ACGT 四种碱基的数量。

V. 平均胎儿浓度和平均测序错误率分别记为 ε 和 e。

第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵：

I. 初始状态概率分布记为 $\pi = \{\pi_j\}$ （ $j \in Q$ ）。

根据本发明的实施例，在没有参考数据的情况下，可以设 $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$，即每种隐藏状态在第一个位点出现的可能性相等。

II. 根据本发明的实施例， 记单倍体重组概率为 $p_r = re/N$，其中 re 代表人类配子基因组重组平均次数，为先验数据在 25 到 30 之间。

III. 根据本发明的实施例，记单倍体重组转移矩阵记为 $A = \{a_{jk}\}$ （ $j,k \in Q$ ），其中 a_{jk} 为隐藏状态转移的概率，即

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r) \cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

胎儿单倍型 $h_0 = \{m_{1,x_1}, ..., m_{i,x_i}, ..., m_{N,x_N}\}$， $h_1 = \{f_{1,y_1}, ..., f_{i,y_i}, ..., f_{N,y_N}\}$ 的下标 x_i 和 y_i 组成的序列对， $q_i = \{x_i, y_i\}$ 组成了我们需要解码的隐藏状态。举个例子， $x_i = 0$ 代表，“母源性染色体上，对应基因座上等位基因型为 $m_{i,0}$ ”。

第二步，构建观察序列概率矩阵：

根据本发明的实施例，记观察序列概率矩阵为 $B = \{b_{i,j}(s_i)\}$ （ $i = 1,2,3,...,N$， $j \in Q$ ），其中 $b_{i,j}(s_i)$ 代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j， $j=\{x_i, y_i\}$ ）时，观测到这种测序信息的可能性”，即

$$b_{i,j}(s_i) = \Pr(s_i | q_i = j, \{m_0, m_1\}) = \frac{(n_{i,A} + n_{i,C} + n_{i,G} + n_{i,T})!}{n_{i,A}! n_{i,C}! n_{i,G}! n_{i,T}!} \cdot (P_{i,A})^{n_{i,A}} \cdot (P_{i,C})^{n_{i,C}} \cdot (P_{i,G})^{n_{i,G}} \cdot (P_{i,T})^{n_{i,T}},$$

其中$P_{i,base}$代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j，$j=\{x_i, y_i\}$）时，该碱基出现的可能性”，即

$$P_{i,base} = \Pr\left(base \middle| q_i = j, \{m_0, m_1\}\right) = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon\cdot\Delta\left(base, f_{y_i}\right),$$

其中指示函数

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$$

。

此步是进行 HMM 参数，每个位点的观察序列概率分布$b_{i,j}\left(s_i\right)$计算，即计算每个位点上不同胎儿单倍型（隐藏状态）下，血浆出现当前测序数据（观察序列）的可能性。

第三步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）:

定义 局部概率 $\delta_i\left(q_i\right) = \left(\max_{q_{i-1}\in Q}\delta_i\left(q_i\right)\cdot a_{q_{i-1}q_i}\right)\cdot b_{i,q_i}\left(s_i\right)$

定义 逆向指针 $\Psi_i\left(q_i\right) = \arg\max_{q_{i-1}\in Q}\delta_i\left(q_i\right)\cdot a_{q_{i-1}q_i}$

这里所使用的术语局部概率$\delta_i\left(q_i\right)$和逆向指针$\Psi_i\left(q_i\right)$都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE, Vol.77，No.2，1989 年 2 月，通过参照将其全文并入本文。

第四步，确定最终状态，并回溯最优路径:

确定最终状态，$q_N^* = \arg\max_{q_N\in Q}\delta_N\left(q_N\right)$

按照逆向指针回溯最优路径，即最可能胎儿单基因型 $q_i^* = \Psi_i\left(q_i\right)$ $(i = 1,2,3,...,N-1)$。

第五步，输出结果。

由此，能够有效地对胎儿基因组的序列进行分析。相比其他已有的产前检测技术方法，本方法有以下技术优势，主要体现在准确性和可获得的遗传信息量上:

1) 根据本发明实施例，检测的位点不仅针对父源性位点，对母源性位点，即母亲杂合位点，同样可以很好地检测出来胎儿是否遗传了母亲的致病位点，检测准确率可高达 95%以上，且可以检测多种变异类型，扩大了疾病检测的范围。

2) 根据本发明实施例，不但可以通过一次测序获得多个位点、多种疾病的信息，对于一些在孕妇血浆中覆盖程度较低，单纯提高测序深度依然无法准确判定的基因序列，可以通过本方法推断得到，结果准确可靠。

3) 根据本发明实施例，可进行遗传疾病作图，对于一些连锁相关疾病，可通过其他位

点的信息直接推断出来，一次可获得的信息量大，对临床检测更加具有指导意义。

另外，根据本发明实施例，本发明的确定胎儿基因组中预定区域碱基信息的方法，不仅限于 SNP 或者 STR 等某一种遗传多态性位点，对所有的遗传多态性位点均可适用，且可以多种位点同时使用，以便互相验证。除了可进行产前无创检测胎儿基因组信息，达到疾病检测的目的，还可以进行无创产前亲子鉴定，在孩子出生前判定孩子父亲身份，为一些涉及抚养责任和义务、财产纠纷、性侵案等协助侦破。

用于确定胎儿基因组中预定区域碱基信息的系统

在本发明的又一方面，本发明提出了一种用于确定胎儿基因组中预定区域核酸序列的系统。根据本发明的实施例，参考图 2，该系统 1000 可以包括：文库构建装置 100、测序装置 200 以及分析装置 400。

根据本发明的实施例，文库构建装置 100 适于针对胎儿基因组 DNA 样本，构建测序文库。根据本发明的实施例，测序装置 200 与文库构建装置 100 相连，并且适于对所构建的测序文库进行测序，以便获得测序结果，所得到的测序结果由多个测序数据构成。根据本发明的实施例，还可以进一步包括 DNA 样本分离装置，该 DNA 样本分离装置适于从孕妇外周血中提取胎儿基因组 DNA 样本。由此，该系统可以适用于进行无创产前检测。

根据本发明的实施例，任选地，还可以包括比对装置 300。根据本发明的实施例，比对装置 300 与测序装置 200 相连，并且适于将所得到的测序结果与参照序列进行比对，以便确定与预定区域对应的测序数据。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46）包括但不限于 Illumina-Solexa（GA™,HiSeq2000™ 等）、ABI-Solid 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing），Pacific Biosciences 公司单分子实时测序（single molecule real-time (SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245）。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。根据本发明的实施例，可以使用的参照序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参照序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI 36.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比

对。

根据本发明的实施例，分析装置 400 适于基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

根据本发明的实施例，惠特比算法采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re=25~30，优选 25，N 为所述预定区域的长度，

采用

$$a_{jk}=\Pr\left(q_i=k|q_{i-1}=j\right)=\begin{cases}\left(1-p_r\right)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ \left(1-p_r\right)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\quad or\quad x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

作为重组转移矩阵， p_r=re/N。

根据本发明的实施例，将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据进一步包括按照下列公式确定概率最高的碱基:

$$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta\left(base,m_k\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,m_{x_i}\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,f_{y_i}\right)$$

其中，

$$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胎儿基因组中预定区域核酸序列的系统。不再赘述。

由此，利用该系统，能够有效地实施前面所述的确定胎儿基因组中预定区域核酸序列的方法，可以通过例如惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

在本文中所述的术语“相连”应作广义理解，既可以是直接相连，也可以是间接相连，只要能够实现上述功能上的衔接即可。

需要说明的是，本领域技术人员能够理解，在前面所描述的确定胎儿基因组中预定区域核酸序列的方法的特征和优点也适合于确定胎儿基因组中预定区域核酸序列的系统，为描述方便，不再详述。

计算机可读介质

在本发明的又一方面，本发明提出了一种计算机可读介质。根据本发明的实施例，计

算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。由此，利用该计算机可读介质，能够有效地实施前面所述的方法，从而可以通过例如惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

根据本发明的实施例，指令适于按照惠特比算法，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，在所述惠特比算法中，采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re=25~30，优选 25，N 为所述预定区域的长度，

采用

$$a_{jk}=\Pr\left(q_i=k|q_{i-1}=j\right)=\begin{cases}\left(1-p_r\right)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ \left(1-p_r\right)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\quad or\quad x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

作为重组转移矩阵，其中，p_r=re/N。

根据本发明的实施例，所述指令将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据进一步包括按照下列公式确定概率最高的碱基:

$$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta\left(base,m_k\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,m_{x_i}\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,f_{y_i}\right)$$

其中，

$$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胎儿基因组中预定区域核酸序列的系统。不再赘述。

此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

就本说明书而言，“计算机可读介质”可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例（非穷尽性列表）包括以下：具有一个或多个布线的电连接部（电子装置），便携式计算机盘盒（磁装置），随机存取存储器（RAM），只读存储器（ROM），可擦除可编辑只读存储器（EPROM 或闪速存储器），光纤装置，以及便携式光盘只读存储器（CDROM）。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为例如可以通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机

存储器中。

应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本发明各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件（例如参考 J.萨姆布鲁克等著，黄培堂等译的《分子克隆实验指南》，第三版，科学出版社）或者按照产品说明书进行。所用试剂或仪器未注明生产厂商者，均为可以通过市购获得的常规产品，例如可以采购自 Illumina 公司。

一般方法

本发明实施例的主要步骤包括:

1) 无创采取含有胎儿遗传物质的孕妇样品，提取其中含有的遗传物质。
2) 胎儿家庭成员如父母和外祖父母等基因组 DNA 提取和纯化。
3) 各遗传物质根据不同测序平台测序要求进行文库构建。
4) 测序获得的数据进行过滤，过滤条件根据质量值、接头污染等来设定。
5) 获得的高质量序列根据需要进行组装处理，组装结果与人类基因组参考序列进行比对。获得唯一比对的序列，带入模型进行分析。

分析模型:

记号:

I.需要检测的位点数为 N。

II. 父母的单倍型分别记为 $FH=\{fh_0,fh_1\}$ 和 $MH=\{mh_0,mh_1\}$，

其中，

$mh_k = \{m_{1,k},...,m_{i,k},...,m_{N,k}\}$， $fh_k = \{f_{1,k},...,f_{i,k},...,f_{N,k}\}$，

$\forall fh_{i,k}, mh_{i,k} \in \{A,C,G,T\}$，

$k \in \{0,1\}$， $i = 1,2,3,...,N$。

III. 将未知的胎儿单倍型记为 $H = \{h_0, h_1\}$，特别地， h_0和 h_1分别遗传自母亲和父亲。

$h_0 = \{m_{1,x_1},...,m_{i,x_i},...,m_{N,x_N}\}$， $h_1 = \{f_{1,y_1},...,f_{i,y_i},...,f_{N,y_N}\}$，

其中， $x_i \in \{0,1\}$， $y_i \in \{0,1\}$，

下标 x_i和 y_i组成的序列对， $q_i = \{x_i, y_i\}$组成了我们需要解码的隐藏状态，

而所有可能出现的隐藏状态组成了集合 Q。

IV. 测序数据记为， $S = \{s_1,...,s_i,...,s_N\}$

其中， $s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$，代表此位点的测序信息，包含了 ACGT 四种碱基的数量。

V. 平均胎儿浓度和平均测序错误率分别记为 ε和 e。

第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵：

I. 初始状态概率分布记为 $\pi = \{\pi_j\}$（ $j \in Q$），

根据本发明的实施例，在没有参考数据的情况下，可以设 $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$，即每种隐藏状态在第一个位点出现的可能性相等。

II. 根据本发明的实施例， 记单倍体重组概率为 $p_r = re/N$，其中 re代表人类配子基因组重组平均次数，为先验数据在 25 到 30 之间。

III. 根据本发明的实施例，记单倍体重组转移矩阵记为 $A = \{a_{jk}\}$（ $j,k \in Q$），其中 a_{jk}为隐藏状态转移的概率，即

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

第二步，构建观察序列概率矩阵：

根据本发明的实施例，记观察序列概率矩阵为 $B = \{b_{i,j}(s_i)\}$（ $i = 1,2,3,...,N$， $j \in Q$），其中 $b_{i,j}(s_i)$代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j）时，观测到这种测序

信息的可能性”，即

$$b_{i,j}(s_i) = \Pr\left(s_i \middle| q_i = j, \{m_0, m_1\}\right) = \frac{(n_{i,A}+n_{i,C}+n_{i,G}+n_{i,T})!}{n_{i,A}!n_{i,C}!n_{i,G}!n_{i,T}!} \cdot \left(P_{i,A}\right)^{n_{i,A}} \cdot \left(P_{i,C}\right)^{n_{i,C}} \cdot \left(P_{i,G}\right)^{n_{i,G}} \cdot \left(P_{i,T}\right)^{n_{i,T}}$$

其中 $P_{i,base}$ 代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j）时，该碱基出现的可能性”，即

$$P_{i,base} = \Pr\left(base \middle| q_i = j, \{m_0, m_1\}\right) = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta\left(base, m_k\right) + \frac{1}{2}\varepsilon \cdot \Delta\left(base, m_{x_i}\right) + \frac{1}{2}\varepsilon \cdot \Delta\left(base, f_{y_i}\right),$$

其中指示函数

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}。$$

第三步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）:

定义 局部概率 $\delta_i(q_i) = \left(\max_{q_{i-1}\in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}\right) \cdot b_{i,q_i}(s_i)$

定义 逆向指针 $\Psi_i(q_i) = \arg\max_{q_{i-1}\in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}$

第四步，确定最终状态状态，并回溯最优路径:

确定最终状态， $q_N^* = \arg\max_{q_N \in Q} \delta_N(q_N)$，

根据逆向指针回溯最优路径，即最可能胎儿单基因型 $q_i^* = \Psi_i(q_i)$ $(i = 1,2,3,...,N-1)$。

第五步，输出结果。

实施例 1

样品收集及处理:

（1）所取样品包括一个家庭中父亲和母亲孕期的外周血，胎儿出生后取脐带血，以EDTA 抗凝管收集，Oragene® DNA 唾液采集/DNA 纯化试剂盒 OG-250 采集祖父母和外祖父母唾液。

（2）祖父母和外祖父母唾液 DNA 提取后用 Infinium® HD Human610-Quad BeadChip 基因芯片进行分型。

（3）取母亲孕期外周血，1600g，4℃离心 10 分钟，将血细胞和血浆分开，血浆再以 16000g，4℃离心 10 分钟，进一步去除残留的白细胞。孕妇血浆用 TIANamp Micro DNA Kit（TIANGEN）提取 DNA，得到母亲和胎儿基因组 DNA 混合物，并且从全血分离血浆后剩余的白细胞中提取母亲基因组 DNA。将所得到的血浆 DNA 根据 Illumia®公司 HiSeq2000™

测序仪的上机要求进行建库，构建好的文库经 Agilent®Bioanalyzer 2100 检测片段分布范围符合要求，再经过 Q-PCR 方法对两个文库进行定量，合格后 Illumina® HiSeq2000™ 测序仪测序，测序循环数为 PE101index（即双向 101bp index 测序），其中仪器的参数设置及操作方法都按照 Illumina®操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

（4）父亲外周血、母亲外周血白细胞和胎儿脐带血则直接用 TIANamp Micro DNA Kit（TIANGEN）提取试剂盒提取基因组 DNA。

除血浆 DNA 样品外，将所获得的所有 DNA 样品，需用 Covaris™ 打断仪打断至 500bp 大小的片段。将获得的 DNA 片段以及血浆 DNA 样品根据 Illumia®公司 HiSeq2000™ 测序仪的上机要求进行建库，具体步骤如下：

末端修复反应体系：

10 × T4 多核苷酸激酶缓冲液	10 μl
dNTPs(10mM)	4 μl
T4 DNA 聚合酶	5 μl
Klenow 片段	1 μl
T4 多核苷酸激酶	5 μl
DNA 片段	30μl

ddH_2O 补齐至 100 μl

20℃反应 30 分钟后，使用 PCR Purification Kit(QIAGEN)回收末端修复产物。将所得到的产物最后溶于 34μl 的 EB 缓冲液中。

末端添加碱基 A 反应体系：

10 × Klenow 缓冲液	5μl
dATP(1mM)	10μl
Klenow (3'-5' exo^-)	3μl
DNA	32μl

37℃温育 30 分钟后，经 MinElute® PCR Purification Kit(QIAGEN)纯化并溶于 12μl 的 EB 中。

接头连接反应体系：

2x Rapid DNA 连接缓冲液	25μl
PEI Adapter oligomix(20uM)	10μl
T4 DNA 连接酶	5μl
添加碱基 A 的产物	10μl

20℃反应 15 分钟后，使用 PCR Purification Kit(QIAGEN)回收连接产物。将所得到的产物最后溶于 32μl 的 EB 缓冲液中。

PCR 反应体系：

接头连接反应产物	10 μl

Phusion DNA Polymerase Mix	25 μl
PCR 引物 (10 pmol/μl)	1 μl
Index N(10 pmol/μl)	1 μl
超纯水	13 μl

反应程序如下:

98℃	30 s	
98℃	10 s	10 个循环
65℃	30 s	
72℃	30 s	
72℃	5 min	
4℃	Hold	

使用 PCR Purification Kit(QIAGEN)回收 PCR 产物。样品最后溶于 50μl 的 EB 缓冲液中。

构建好的文库经 Agilent®Bioanalyzer 2100 检测片段分布范围符合要求，再经过 Q-PCR 方法对两个文库进行定量，合格后，用 Illumina® HiSeq2000™ 测序仪测序，测序循环数为 PE101index（即双向 101bp index 测序），其中仪器的参数设置及操作方法都按照 Illumina® 操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

（5）父母基因组测序分型:

a. 使用 SOAP2 将测序数据比对到人类参考基因组（版本为 NCBI 36.3，HG18）。

b. 使用 SOAPsnp（南方汉族（CHS）家系数据使用的是千人计划数据）对数据进行一致序列（consensus sequence，CNS）构建。

c. 提取出标记位点的基因型。

（6）父母单倍体型推断:

a.构建含祖辈与父母基因型的群体基因型矩阵，即提取父母、祖辈和南方汉族家系在标记位点的基因型。

b. 使用 BEAGLE 对父母的单倍型进行推断。

（7）胎儿单倍体型推断:

a. 用 SOAP2 将血浆测序数据比对到人类参考基因组（版本为 NCBI 36.3，HG18）。

b. 构建初始状态概率向量，以及单倍体重组转移矩阵。

构建初始状态概率向量：采取无参考数据模式，即各个初始状态概率相等，均为 0.25。

单倍体重组转移矩阵：保守地，我们取 $re=25$（其余按一般方法所述）。

c. 统计每个位点的测序信息，并构建观察序列概率矩阵（其余按一般方法所述）。

d. 构建局部概率矩阵，和逆向指针（其余按一般方法所述）。

e. 确定最终状态状态，并回溯最优路径。

f. 输出。

根据胎儿出生后的脐带血基因分型结果，我们的分类准确性统计如下:

			母亲						合计		
			纯合			杂合					
			位点数	准确数	准确率	位点数	准确数	准确率	位点数	准确数	准确率
常染色体	父亲	纯合	199,552	199,552	100.00%	66,238	63,968	96.57%	265,790	263,520	99.15%
		杂合	65,409	64,735	98.97%	41,849	39,944	95.45%	107,258	104,679	97.60%
	合计		264,961	264,287	99.75%	108,087	103,912	96.14%	373,048	368,199	98.70%
X染色体			4,881	4,881	100.00%	1,718	1,478	86.03%	6,599	6,359	96.36%

工业实用性

本发明的确定胎儿基因组中预定区域碱基信息的方法、用于确定胎儿基因组中预定区域碱基信息的系统以及计算机可读介质，能够有效地应用于对胎儿基因组中预定区域的核酸序列进行分析。

尽管本发明的具体实施方式已经得到详细的描述，本领域技术人员将会理解。根据已经公开的所有教导，可以对那些细节进行各种修改和替换，这些改变均在本发明的保护范围之内。本发明的全部范围由所附权利要求及其任何等同物给出。

在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示意性实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

权利要求书

1、一种确定胎儿基因组中预定区域碱基信息的方法，其特征在于，包括下列步骤：

针对胎儿基因组 DNA 样本，构建测序文库；

对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；以及

基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

2、根据权利要求 1 所述的方法，其特征在于，所述胎儿基因组 DNA 样本是从孕妇外周血中提取的。

3、根据权利要求 1 所述的方法，其特征在于，所述测序是利用选自 Illumina-Solexa、ABI-Solid、Roche-454 和单分子测序装置的至少一种对所述测序文库进行的。

4、根据权利要求 1 所述的方法，其特征在于，进一步包括将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果。

5、根据权利要求 4 所述的方法，其特征在于，所述参照序列为人类参考基因组。

6、根据权利要求 1 所述的方法，其特征在于，所述胎儿遗传相关个体是所述胎儿的父母。

7、根据权利要求 1 所述的方法，其特征在于，按照惠特比算法，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

8、根据权利要求 7 所述的方法，其特征在于，在所述惠特比算法中，采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re=25~30，优选 25，N 为所述预定区域的长度，

采用

$$a_{jk}=\Pr\left(q_i=k\middle|q_{i-1}=j\right)=\begin{cases}\left(1-p_r\right)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ \left(1-p_r\right)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\quad or\quad x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

作为重组转移矩阵，其中，p_r=re/N。

9、根据权利要求 4 所述的方法，其特征在于，将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果进一步包括按照下列公式确定概率最高的碱基：

$$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta\left(base,m_k\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,m_{x_i}\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,f_{y_i}\right)$$

其中，

$$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

10、根据权利要求 1 所述的方法，其特征在于，进一步包括：所述预定区域为已知存在遗传多态性的位点。

11、根据权利要求 10 所述的方法，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

12、一种用于确定胎儿基因组中预定区域碱基信息的系统，其特征在于，包括：

文库构建装置，所述文库构建装置适于针对胎儿基因组 DNA 样本，构建测序文库；

测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；

分析装置，基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

13、根据权利要求 12 所述的系统，其特征在于，进一步包括 DNA 样本分离装置，所述 DNA 样本分离装置适于从孕妇外周血中提取胎儿基因组 DNA 样本。

14、根据权利要求 12 所述的系统，其特征在于所述测序装置为选自 Illumina-Solexa、ABI-Solid、Roche-454 和单分子测序装置的至少一种。

15、根据权利要求 12 所述的系统，其特征在于，进一步包括比对装置，所述比对装置与所述测序装置相连，用于将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果。

16、根据权利要求 12 所述的系统，其特征在于，所述分析装置适于按照惠特比算法，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

17、根据权利要求 16 所述的系统，其特征在于，所述惠特比算法采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re=25~30，优选 25，N 为所述预定区域的长度，

采用

$$a_{jk}=\Pr(q_i=k|q_{i-1}=j)=\begin{cases}(1-p_r)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ (1-p_r)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\quad or\quad x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

作为重组转移矩阵，其中，p_r=re/N。

18、根据权利要求 15 所述的系统，其特征在于，将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果进一步包括按照下列公式确定概率最高的碱基：

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta(base, m_k) + \frac{1}{2}\varepsilon\cdot\Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon\cdot\Delta(base, f_{y_i})$$

其中，

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$$

。

19、一种计算机可读介质，其特征在于，所述计算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定预定区域的碱基信息。

20、根据权利要求 19 所述的计算机可读介质，其特征在于，所述指令适于按照惠特比算法，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

21、根据权利要求 20 所述的计算机可读介质，其特征在于，在所述惠特比算法中，采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re=25~30，优选 25，N 为所述预定区域的长度，

采用

$$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

作为重组转移矩阵，其中，p_r=re/N。

22、根据权利要求 19 所述的计算机可读介质，其特征在于，所述指令将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果。

23、根据权利要求 22 所述的计算机可读介质，其特征在于，所述指令将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果进一步包括按照下列公式确定概率最高的碱基:

$$P_{i,base} = \sum_{k\in\{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta(base, m_k) + \frac{1}{2}\varepsilon\cdot\Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon\cdot\Delta(base, f_{y_i})$$

其中，

$$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}$$

。

24、根据权利要求 19 所述的计算机可读介质，其特征在于，进一步包括：所述预定区域为已知存在遗传多态性的位点。

25、根据权利要求 24 所述的计算机可读介质，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

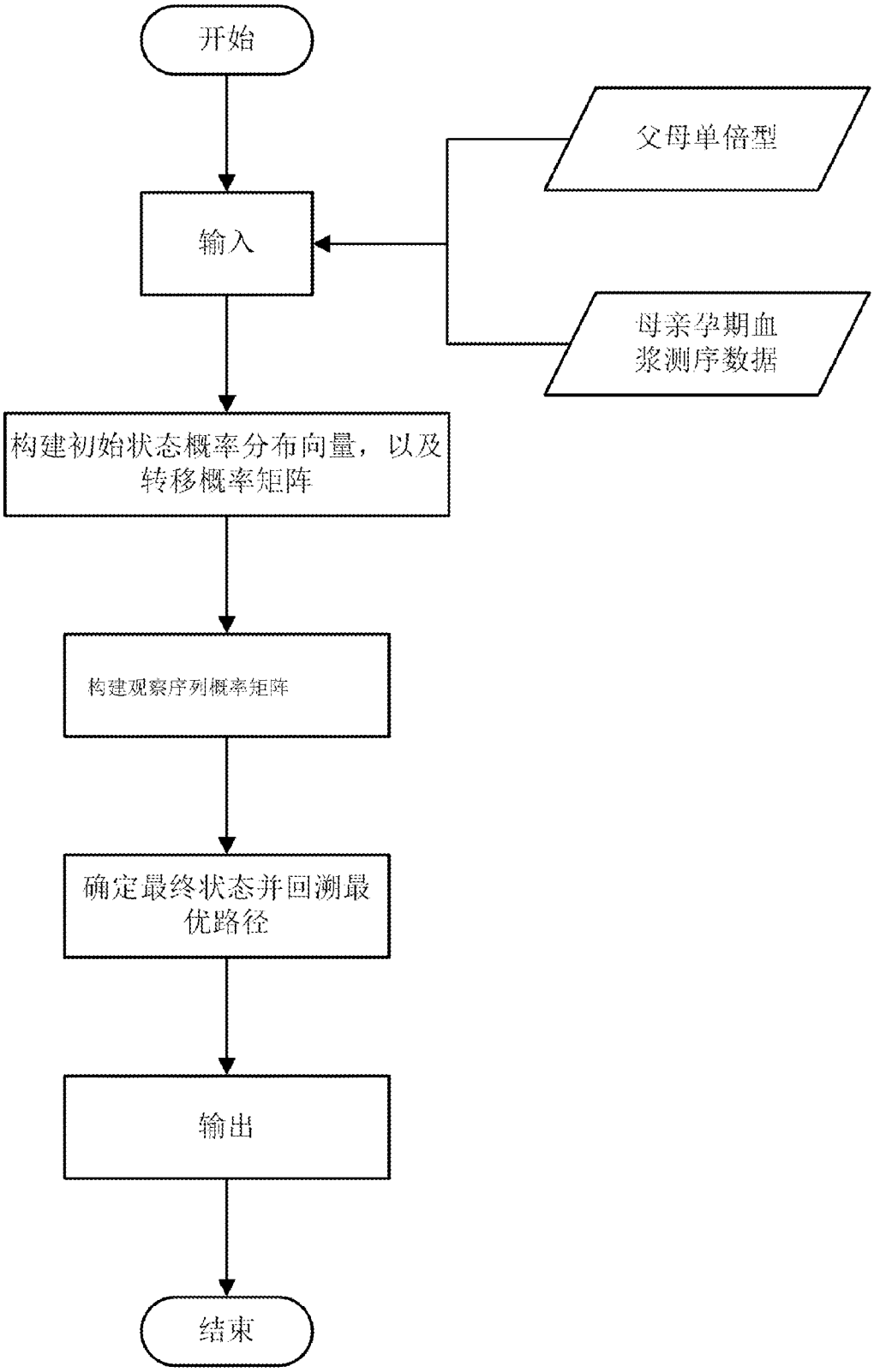

开始
父母单倍型
输入
母亲孕期血浆测序数据
构建初始状态概率分布向量，以及转移概率矩阵
构建观察序列概率矩阵
确定最终状态并回溯最优路径
输出
结束

图 1

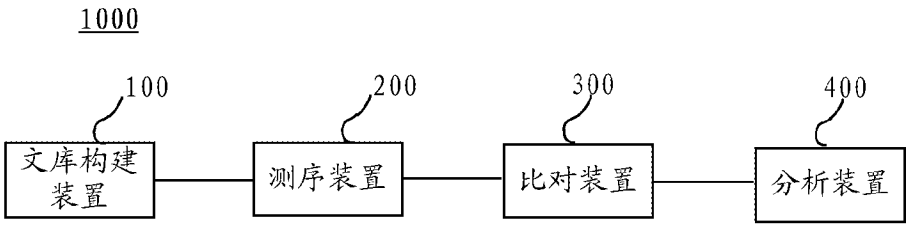


图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/075478

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i

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)

IPC: 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)

CNABS, CPRSABS, SIPOABS, DWPI, CNTXT, WOTXT, EPTXT, USTXT, CNKI, GOOGLE SCHOLAR, Elsevier Science, PubMed: hidden Markov model, HMM, viterbi algorithm, library, sequencing, fetus, foetus, fetal, foetal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 102127818 A (ZHANG, Kang) 20 July 2011 (20.07.2011) , claims 1-7, description, page 3, paragraph [0035] to page 4, paragraph [0050]	1-18
Y	CN 102317473 A (PACIFIC BIOSCIENCES CALIFORNIA INC.) 11 January 2012 (11.01.2012) , claims 1 and 2, description, page 32, paragraph [0003] to page 33, paragraph [0001]	1-18

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

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

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

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

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

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

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

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

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

"&"document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 January 2013 (27.01.2013)	14 February 2013 (14.02.2013)
Name and mailing address of the ISA State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	Authorized officer ZHAO, Yanhao Telephone No. (86-10) 62411043

Form PCT/ISA /210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2012/075478

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.:19-25
because they relate to subject matter not required to be searched by this Authority, namely:
Claims 19-25 are directed to mere presentations of information (PCT Rule 39. 1(v)).

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2012/075478

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102127818 A	20.07.2011	None	
CN 102317473 A	11.01.2012	WO 2010/068289 A2	17.06.2010
		US 2011/183320 A1	28.07.2011
		US 2010/221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010/068289 A3	14.10.2010
		AU 2009/325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011

Form PCT/ISA /210 (patent family annex) (July 2009)

国际检索报告

国际申请号
PCT/CN2012/075478

A. 主题的分类

C12Q 1/68 (2006.01) i

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC: C12Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNABS,CPRSABS,SIPOABS,DWPI,CNTXT,WOTXT,EPTXT,USTXT,CNKI,GOOGLE SCHOLAR, Elsevier Science, PubMed：隐马尔可夫模型，惠特比算法，文库，测序，隐马可夫, 隐马氏，隐 MarkoV, 胎儿，hidden Markov model,HMM, viterbi algorithm, library, sequencing, fetus, foetus, fetal, foetal

C. 相关文件

类　型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	CN102127818 A (张康) 20.7 月 2011（20.07.2011）,权利要求 1-7，说明书第 3 页第 35 段-说明书第 4 页第 50 段	1-18
Y	CN102317473 A (加利福尼亚太平洋生物科学股份有限公司) 11.1 月 2012（11.01.2012）,权利要求 1-2，说明书第 32 页第 3 段-第 33 页第 1 段	1-18

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型:

"A" 认为不特别相关的表示了现有技术一般状态的文件

"E" 在国际申请日的当天或之后公布的在先申请或专利

"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

"O" 涉及口头公开、使用、展览或其他方式公开的文件

"P" 公布日先于国际申请日但迟于所要求的优先权日的文件

"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
27.1 月 2013(27.01.2013)	**14.2 月 2013 (14.02.2013)**
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号：　(86-10)62019451	受权官员 赵彦豪 电话号码：(86-10) **62411043**

PCT/ISA/210 表(第 2 页) (2009 年 7 月)

国际检索报告

国际申请号
PCT/CN2012/075478

第II栏　某些权利要求被认为是不能检索的意见(续第 1 页第 2 项)

根据条约第 17 条(2)(a)，对某些权利要求未做国际检索报告的理由如下：

1. ☒ 权利要求：19-25

因为它们涉及不要求本单位进行检索的主题，即：

权利要求 19-25 涉及单纯的信息提供（PCT 细则 39.1 (v)）。

2. ☐ 权利要求：

因为它们涉及国际申请中不符合规定的要求的部分，以致不能进行任何有意义的国际检索，

具体地说：

3. ☐ 权利要求：

因为它们是从属权利要求，并且没有按照细则 6.4(a)第 2 句和第 3 句的要求撰写。

第III栏　缺乏发明单一性的意见(续第 1 页第 3 项)

本国际检索单位在该国际申请中发现多项发明，即：

1. ☐ 由于申请人按时缴纳了被要求缴纳的全部附加检索费，本国际检索报告涉及全部可作检索的权利要求。

2. ☐ 由于无需付出有理由要求附加费的劳动即能对全部可检索的权利要求进行检索，本单位未通知缴纳任何附加费。

3. ☐ 由于申请人仅按时缴纳了部分被要求缴纳的附加检索费，本国际检索报告仅涉及已缴费的那些权利要求。

具体地说，是权利要求：

4. ☐ 申请人未按时缴纳被要求缴纳的附加检索费。因此，本国际检索报告仅涉及权利要求书中首先提及的发明；包含该发明的权利要求是：

关于异议的说明： ☐ 申请人缴纳了附加检索费，同时提交了异议书，适用时，缴纳了异议费。

☐ 申请人缴纳了附加检索费，同时提交了异议书，但未在通知书规定的时间期限内缴纳异议费。

☐ 缴纳附加检索费时未提交异议书。

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2012/075478

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN102127818 A	20.07.2011	无	
CN 102317473 A	11.01.2012	WO 2010068289 A2	17.06.2010
		US 2011183320 A1	28.07.2011
		US 2010221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010068289 A3	14.10.2010
		AU 2009325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 December 2012 (13.12.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/170429 A1

(51) International Patent Classification:
B65G 17/22 (2006.01) *B65G 47/96* (2006.01)

(21) International Application Number:
PCT/US2012/040926

(22) International Filing Date:
5 June 2012 (05.06.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
13/156,227 8 June 2011 (08.06.2011) US

(71) Applicant *(for all designated States except US)*: **UNITED STATES POSTAL SERVICE** [US/US]; 475 L'Enfant Plaza, #6443, Washington, DC 20260 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **CLOSE, Donald, R.** [US/US]; 831 Glenbrook Road, Fairfax, VA 22031 (US). **BOMBAUGH, Scott, R.** [US/US]; 9350 Burke Road, Burke, VA 22015 (US).

(74) Agent: **HART, Daniel**; Knobbe Martens Olson & Bear Llp, 2040 Main Street, Fourteenth Floor, Irvine, CA 92614 (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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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, 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, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: JOINT FOR CONNECTING CARRIER UNITS

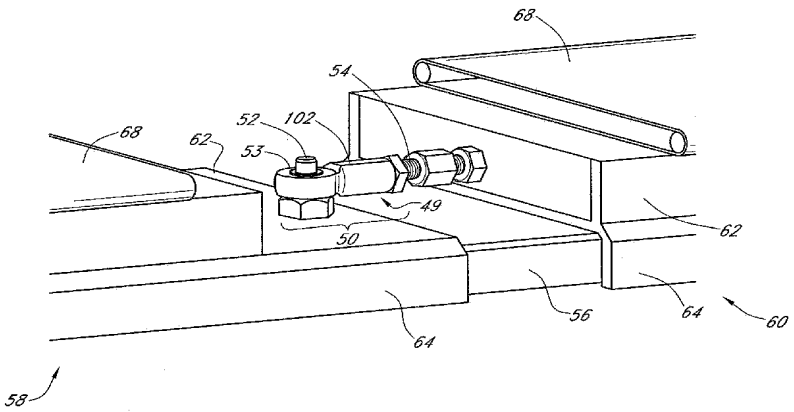


FIG. 2

(57) **Abstract**: The technology disclosed provides a joint (50) for connecting carrier units (58,60) together so as to dissipate and absorb axial forces experienced by the carrier units. The joint (50) may be comprised of a joint housing (102) and an elastomeric insert (112). The housing (102) may be comprised of a body portion (114) and a head portion (103) and the head portion may include an annular flange (104) in which the elastomeric insert (112) is configured to be secured. The elastomeric joint (50) is capable of absorbing and dissipating horizontal, rotational, and vertical forces experienced by the carrier units (58,60) in non-linear travel along a track(56).

WO 2012/170429 A1

JOINT FOR CONNECTING CARRIER UNITS

BACKGROUND OF THE INVENTION

Field of the Development

[0001] This application relates generally to conveyor systems for transporting objects, and more specifically refers to methods and devices for dissipating forces in conveyor joints.

Description of the Related Art and Problem to Be Solved

[0002] Conveyor systems are useful in selecting, sorting, and transporting objects from one location to another. A conveyor system may be made up of individual conveyor units, or "carriers," which are connected together by joints to make up a conveyor train. The conveyor train typically rides on a track in a continuous loop. An individual carrier may be equipped with a discharge device such as a motor-driven cross-belt, a tilt-tray, or a diverter paddle, etc. for discharging objects carried atop the carrier in a direction transverse to the direction traveled by the conveyor train. The discharge device is programmed to activate at a pre-programmed time and location along the conveyor track based on, for example, the characteristics of the object being transported.

[0003] A conveyor system may be employed in package sorting facilities, and as such will often travel a circuitous route through the facility, receiving and distributing packages at multiple locations. In the process, a conveyor system may be required to travel through multiple curves of varying radii and additionally travel up and down inclines of varying elevations. Non-linear travel in a conveyor system creates difficulty in establishing and maintaining a constant speed for the conveyor train. As a result, the non-linear travel may also cause timing irregularities for the discharge devices in discharging the objects being carried.

[0004] Also, conveyor systems with curved tracks and elevation changes tend to experience significant mechanical failure at a higher frequency than do conveyor systems with straight tracks alone. Such mechanical failure necessitates the need for frequent stoppage of the system for maintenance and repair. The joints connecting the conveyor carriers are particularly vulnerable to flexural fatigue in non-linear conveyor systems. The joints provide a point of connection between the conveyor carriers and thus facilitate turning of the conveyor train through the curves. Much of the axial forces experienced by the conveyor system are, thus, focused on the joints. The axial forces may cause flexural fatigue in the joints, leading to

deterioration of the structural integrity of the metal joint components. When the joints deteriorate, many of the joint components may become loose or fit poorly together and require frequent repair. Joint deterioration may also cause the joint components to repeatedly collide together, causing high-decibel noise levels unsuitable for workers attending the conveyor system.

[0005] An inability of the joints to absorb the axial forces experienced in a non-linear conveyor system leads not only to break down of the joints themselves, but to transmission of the axial forces from the joints into the frame of the conveyor system. The transmitted forces may cause damage to the track and conveyor system support structure. In this manner, a non-linear conveyor system has historically required longer and more frequent maintenance stoppage time, resulting in costly repairs to the conveyor track and conveyor carriers, not to mention the opportunity cost of not being able to run the conveyor system as long as one would ideally like.

[0006] Previous attempts at mitigating the forces exhibited in non-linear conveyor systems have included the use of slip joints disposed between conveyor carriers at periodic locations along the carrier train. For practical reasons, the slip joints cannot be used between each individual conveyor carrier without causing slack to build up in the conveyor train. Thus, the forces experienced by the majority of individual joints connecting the conveyor carriers are not dissipated even with the use of slip joints.

[0007] Another attempt at mitigating the forces in non-linear conveyor systems has included the use of ball joints to connect the conveyor carriers. Often Teflon or other coating materials have been placed over the ball of the ball joint or on the inside of the housing to protect the joint from frictional forces that the joint experiences in a non-linear conveyor system. The use of a ball joint has proven inadequate, however, in dealing with the forces experienced in the non-linear conveyor system because the ball joint is unable to dissipate the axial forces experienced by the joint. The Teflon coating simply acts as a small buffer between the housing and ball portions of the ball joint to prevent wear on the ball. The joint is unable to dissipate and control the axial forces experienced by the joint in a curve in a non-linear conveyor system. The coating material is liable to wear out quickly and need frequent replacement. Furthermore, the axial forces experienced by the joints are still transferred to the overall conveyor system, leading to system damage and potential timing irregularities.

SUMMARY OF THE INVENTION

[0008] Some embodiments of the disclosure provide for a joint for connecting carrier units, including a joint housing and an elastomeric insert. The joint housing includes a head portion on a first side comprising an annular flange. The annular flange includes interior walls

that define a hole and a counter bore. The joint housing also includes an elongated body on a second side, wherein the body comprises a bore extending longitudinally therethrough. The elastomeric insert can be positioned inside the counter bore of the annular flange.

[0009] The cross-sectional diameter of the counter bore can be greater than the diameter of the hole in the annular flange. The elastomeric insert can include an outer ring and an inner ring, wherein the inner and outer rings are connected together by supports. The joint may further include a front and a rear gap in the elastomeric insert, each gap defined by the inner walls of the outer ring and the outer walls of the inner ring. The elastomeric insert may include a bumper extending from the inner elastomeric ring and positioned in one of the front or rear gaps and the elastomeric insert may include an elastomeric knob extending from an outer surface of the outer ring. The elastomeric insert can be configured in the counter bore to orient the bumper towards a front side of the annular flange and the knob can be configured to fit into a connecting lumen in a rear interior wall of the counter bore.

[0010] Some embodiments of the disclosure provide a bearing, the bearing having a base portion and a tubular portion extending from the base portion, wherein the bearing is configured to be positioned inside the counter bore of the annular flange and the elastomeric element configured to fit around the tubular portion of the bearing. A bottom portion of the elastomeric insert can be secured to the base portion of the bearing and a top portion of the elastomeric insert can be secured to a top of the counter bore of the annular flange.

[0011] Some embodiments of the disclosure provide a method of dissipating forces between carrier units by connecting a first carrier unit to a second carrier unit using an elastomeric joint, receiving a force on the carrier units, and dissipating the force in the elastomeric joint. The elastomeric joint can include a joint housing and an elastomeric insert. The step of connecting the first carrier unit to the second carrier unit can include connecting a first post on the first carrier to an annular flange on the elastomeric joint and connecting a second post on the second carrier to a longitudinal bore in the elastomeric joint.

[0012] The step of dissipating the force in the elastomeric joint can further include receiving the force experienced by the first carrier at the first post at the elastomeric insert and the elastomeric insert is configured to dissipate the force by elastically adjusting within the joint housing and by absorbing the force.

[0013] The method of dissipating forces between carrier units can further include connecting a plurality of carrier units using a plurality of elastomeric joints, wherein the plurality of carrier units comprise a conveyor train. The step of receiving a force can be a result of the

plurality of carrier units traveling in a curve. The step of receiving a force can be a result of the plurality of carrier units traveling on an incline.

[0014] Some embodiments of the disclosure provide a joint for dissipating forces in carrier units including means for connecting conveyor units, means for receiving a force on the conveyor units, means for dissipating the force within the connecting means. The connecting means can be an elastomeric joint and the dissipating means can be elastomeric insert.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015] Fig. 1 is top plan view of certain components of a conveyor system, in which a conveyor train made up of multiple conveyor carriers is traveling around a curve in the conveyor track.

[0016] Fig. 2 is a perspective view of one embodiment of an elastomeric joint connecting two conveyor carriers.

[0017] Fig. 3A is a bottom plan view of an embodiment of the joint housing.

[0018] Fig. 3B is a cross-sectional side view of the joint housing taken along line 3B of Fig. 3A.

[0019] Fig. 4A is a top plan view of one embodiment of an elastomeric insert.

[0020] Fig. 4B is cross-sectional side view of the elastomeric insert taken along line 4B of Fig. 4A.

[0021] Fig. 4C is cross-sectional side view of the elastomeric insert taken along line 4C of Fig. 4A.

[0022] Fig. 5A is a bottom plan view of an embodiment of an elastomeric joint.

[0023] Fig. 5B is a cross-sectional side view of the elastomeric joint taken along line 5B of Fig. 5A.

[0024] Fig. 6 is a perspective view of an embodiment of an elastomeric insert.

[0025] Fig. 7 is a perspective view of an embodiment of a bearing for use in an elastomeric joint.

[0026] Fig. 8A is bottom plan view of an embodiment of an elastomeric joint.

[0027] Fig. 8B is a cross-sectional side view of the elastomeric joint taken along line 8B of Fig. 8A.

[0028] Fig. 8C is a cross-sectional side view of the elastomeric joint taken along line 8C of Fig. 8A.

DETAILED DESCRIPTION OF EMBODIMENTS OF THE DEVELOPMENT

[0029] The features, aspects and advantages of the present development will now be described with reference to the drawings which are intended to be within the scope of the embodiments disclosed herein. These and other embodiments will become readily apparent to those skilled in the art from the following detailed description of the embodiments having reference to the attached figures, the development not being limited to any particular embodiment(s) herein disclosed.

[0030] The technology described herein provides a joint for connecting conveyor carriers together as a conveyor train. The conveyor train travels on a track along a circulating path as part of an overall conveyor system. In the context of an exemplary embodiment of a sorting facility such as is used in a parcel sorting facility, the conveyor system includes one or more package pick-up locations and one or more package discharge locations. An individual carrier may include a discharge device for discharging packages at a programmed time and location. Based on the characteristics of a particular package (e.g., size, delivery destination, parcel method, etc.), the discharge device of the carrier on which the package is riding will be activated to discharge the package from the carrier at a programmed location along the path of the conveyor train. Thus, the conveyor train will preferably operate at a consistent and accurate speed in order to accurately discharge the packages off the conveyor carriers at the correct location. If the speed of the train becomes inconsistent (speeding up or down) or travels at an actual speed different than the programmed speed, the packages may be discharged at the incorrect location.

[0031] The operational demands of a package sorting facility may require that the conveyor system operate continuously for days or weeks on end, without stoppage time for maintenance. Furthermore, when maintenance is required, the stoppage time should be minimal and the actual maintenance performed should ideally be minimal, not requiring significant repairs.

[0032] The basis for the existence of the forces experienced by the joints connecting the conveyor carriers in a non-linear conveyor system are disclosed herein, as are methods and devices for dissipation of the forces.

[0033] Fig. 1 is a top plan view of certain components of a conveyor system 10, in which a conveyor train 12 made up of multiple conveyor carriers 14 is traveling around a curve in a conveyor track 16. The conveyor carriers 14 are connected together by joints 18. Historically the joints 18 connecting the conveyor carriers 14 have been rigid connections (such as ball joints) or have included periodic slip joints between the conveyor carriers 14. Each

individual carrier 14 is of a certain rectilinear length and thus individually cannot conform to the curvature 20 of the track 16. As a result, as a conveyor carrier travels in the curve, the distance 28 from the center of the curvature 32 to the track curvature 20 is greater than the distance 30 from the center of the curvature 32 to the center midpoint 22 of the individual carrier 14 traveling in the curve. Thus, the front and rear midpoints 24, 26 of the conveyor carrier 14 are at a greater distance from the center of curvature 32 than is the center midpoint 22 of the conveyor carrier 14.

[0034] Being at a greater distance from the center of curvature 32, the front and rear midpoints 24, 26 of the carrier must accelerate relative to the center midpoint 22 of the carrier while traveling through the curve. This relative acceleration at the front and rear midpoints 24, 26 induces substantial forces on the conveyor carriers 14, particularly at the joints 18 near the front and rear midpoints 24, 26. Likewise, as the conveyor carriers 14 exit the curve, the conveyor carriers 14 will then decelerate as they enter the linear portion of the track 16. The forces experienced by the conveyor carriers 14 at the joints 18 are experienced primarily in an axial direction, tangential to the radius of curvature.

[0035] The forces that the conveyor carriers 14 experience, particularly at the joints 18, as they enter and exit the curve, are not isolated just to the conveyor carriers 14 traveling through the curve at that time, but instead are propagated through the train 12 of, what have historically been, rigidly-connected conveyor carriers 14. Thus, even conveyor carriers 14 now on a straight portion of the track 16 still experience the propagating forces. In this manner, travel of the conveyor train 12 through a single curve can be the cause of oscillating horizontal forces propagating through the entire conveyor train 12. The forces propagating through the conveyor train 12 also create high-decibel noise levels in the conveyor system 10, noise levels which are potentially harmful to operations staff.

[0036] Thus, the potential for inaccurate timing and the structural damage experienced in a curved conveyor system 10 are often a result of the oscillating axial forces which result from the acceleration and deceleration experienced by the conveyor carriers 14 as they enter and exit the curve. With an understanding of the distribution of forces within the conveyor system 10, a focus on alleviating the forces is possible. The present disclosure provides devices and methods for mitigating and dissipating the forces experienced by the conveyor carriers 14, allowing for a consistent and accurate speed of the conveyor train and reduced stoppage time and reduced maintenance costs.

[0037] Fig. 2 provides a perspective view of an embodiment of an elastomeric joint 50 disclosed herein. The elastomeric joint 50 can be positioned to connect a front carrier 58 and

a rear carrier 60. The elastomeric joint 50 can be attached to a vertical post 52 located on the conveyor base 62 of the front carrier 58 and to a horizontal post 54 located on a conveyor base 62 of the rear carrier 60. The horizontal post 54 on the rear carrier 60 may be threaded to receive a threaded bore (not shown) on the elastomeric joint 50. Each conveyor carrier 58, 60 may further include a discharge device 68 and a conveyor saddle 64 as part of the carrier base 62. The conveyor saddle 64 fits around a conveyor track 56 on which the conveyor carriers 58, 60 travel.

[0038] The elastomeric joint 50 disclosed herein is comprised of a joint housing 102 and an elastomeric insert 53, which is partially visible in Fig. 2. A bottom portion 49 of elastomeric joint 50 faces downwards towards the conveyor track 56. Fig. 3A is a bottom plan view of an embodiment of the joint housing 102. It is noted that the bottom plan of the joint housing 102, as viewed in Fig. 3A, corresponds to the side of the elastomeric joint 50 facing down towards the track 56 in Fig. 2. The joint housing 102 includes a linear body portion 114 and a rounded head portion 103. The head portion 103 includes an annular flange 104, the annular flange 104 having interior walls 105 that define a longitudinal hole 110. The annular flange 104 also comprises interior walls 107 that define a counter bore 122 beginning on a bottom section 108 of the annular flange 104 and extending through the annular flange 104 to culminate at an annular top 116. The counter bore 122, is of a diameter greater than the diameter of the hole 110 in the annular flange 104. In this manner, the annular top 116 of the counter bore 122 circumscribes the hole 110 in the annular flange 104.

[0039] Fig. 3B is a cross-sectional side view of the joint housing 102 taken along line 3B of Fig. 3A. Visible in Fig. 3B is a bore 118 formed within the joint housing 102 so as to extend longitudinally through the body portion 114 of the joint housing 102. The bore 118 may comprise threads 119 extending along the interior face of the bore 118 in the body portion 114. The threads 119 allow the joint housing 102 to be threadably connected to the horizontal post 54 on the rear carrier 60, which may also be threaded along its mating surface.

[0040] The joint housing may also include a connecting lumen 120 passing from the bore 118 through the body portion 114 and the flange 104 to the counter bore 122. The annular flange 104 comprises a top section 106 and a bottom section 108. The interior walls 105 of the annular flange in the top section 106 define the longitudinal hole 110. The counter bore 122 may be a circular shape about a longitudinal axis 121 in order to fit a circular elastomeric insert discussed below, or may be of any other necessary shape to conform to the shape of the desired elastomeric insert. The counter bore 122 comprises a front interior wall 115 and a back interior wall 117. The back interior wall 117 of the counter bore 122 may include the connecting lumen

120 extending therethrough to the bore 118 in the body portion 114 of the joint housing 102. Shown in Fig. 3B is the annular top 116 to the counter bore 122, circumscribing the hole 110 located in the top section 106 of the annular flange 104. The joint housing 102 may be comprised of stainless steel, aluminum, titanium, or other suitable metal or alloy.

[0041] Fig. 4A provides a top plan view of an embodiment of an elastomeric insert 112 which, when in use, may positioned in the joint housing 102 to comprise the elastomeric joint 50 (Fig. 1). This embodiment of the elastomeric insert 112 comprises a cylindrical outer ring 126 and a cylindrical inner ring 124 extending within and parallel to the outer ring 126. The inner ring comprises a channel 132 extending longitudinally therethrough. The inner and outer rings 124, 126 can be connected by supports 128, which in one embodiment extend on opposite sides between the inner and outer rings 124, 126. The connection of the inner and outer rings 124, 126 by the supports 128 creates semi-circular gaps between the facing walls of those rings to define a front gap 130 and a rear gap 131. Extending from the outside surface of the inner ring 124 in the front gap 130 can be a bumper 136 comprising a raised area such as a shoulder or bump, which may be formed as part of the elastomeric insert 112. The bumper 136 extends a portion of the distance between the inner and outer rings 124, 126 within the front gap 130 and provides structural support to the elastomeric insert 112, particularly when the elastomeric joint 50 is experiencing forces due to elevation changes on the conveyor track as described below.

[0042] The outside surface of the outer ring 126 can also include a knob 134, which may also be formed as part of the elastomeric insert 112, and is used to securely fit the elastomeric insert 112 in the joint housing 102 as described below.

[0043] As shown in Fig. 4A, the cross-sectional width 139 of the outer ring 126 at a front section 138 of the elastomeric insert 112 may be smaller than is the cross-sectional width 137 of the outer ring 126 at a rear section 140. A smaller cross-sectional width 139 for the outer ring 126 at the front section 138 allows for a larger front gap 130 in which to position or form the bumper 136. The elastomeric insert 112 may be made of polyurethane, styrene butadiene, natural rubber, urethane, silicone, neoprene, nitrile butadiene, or other such suitable elastomeric material capable of absorbing and dissipating axial forces.

[0044] Fig. 4B is a cross-sectional view of the elastomeric insert 112 taken along line 4B of Fig. 4A. Shown in Fig. 4B is the channel 132 formed by the inner ring 124. Also shown are the front gap 130 and the rear gap 131 between the inner ring 124 and the outer ring 126. Also shown are the knob 134 and the bumper 136, the knob 134 extending from the outside surface of the outer ring 126 and the bumper extending from the outside surface of the inner ring 124 and positioned in the front gap 130. The bumper 136 may, however, extend from the outside

surface of the inner ring 124 into the rear gap 131, instead of into the front gap 130. Alternatively, the bumper 136 may be configured to extend from the inside surface of the outer ring 126, either in the front or rear gaps 130, 131. The longitudinal length of the inner ring 124 may be configured to be larger than the longitudinal length of the outer ring 126 in order to accommodate placement of the elastomeric insert 112 in the joint housing 102, as discussed below.

[0045] Fig. 4C is a cross-sectional view of the elastomeric insert 112 displayed in Fig. 4A taken along line 4C of Fig. 4A. As seen in Fig. 4C, the outer ring 126 may be attached to the inner ring 124 by the supports 128 on at least two sides. The channel 132 is formed by the inner ring 124.

[0046] Fig. 5A is a bottom plan view of the assembled elastomeric joint 50. The illustrated bottom portion corresponds to the bottom side 49 of the elastomeric joint 50 which is illustrated as facing down towards the track 56 in Fig. 2. The body portion 114 of the joint housing 102 is connected to the annular flange 104 as described in reference to Fig. 3A. The elastomeric insert 112 fits firmly within the counter bore 122 (see Fig. 3B) of the joint housing 102.

[0047] Fig. 5B is a cross-sectional side view of the assembled elastomeric joint 50 of Fig. 5A taken along line 5B of Fig. 5A. Here, the inner ring 124 includes the channel 132 extending longitudinally therethrough and can include the bumper 136 extending from the outside of the inner ring in the front gap 130. The outer ring 126 fits firmly in the counter bore 122 against the interior walls 107 and top 116 of the counter bore 122. Because the longitudinal length of the inner ring 124 may be configured to be larger than the longitudinal length of the outer ring 126, the inner ring 124 may extend into or through the hole 110 in the annular flange 104. The top surface of the inner ring 124 may be made flush with the top surface 106 of the annular flange 104, or may extend to a height greater than the top 106 of the annular flange 104, thereby extending beyond the plane defined by the top 106 of the annular flange 104. The bottom surface of the inner and outer rings 124, 126 of the elastomeric insert 112 can be flush with the bottom surface 108 of the annular flange 104 or may be slightly recessed into the counter bore 122.

[0048] The knob 134 on the outside surface of the outer ring 126 can be fitted into the connecting lumen 120. Positioning the knob 134 in the connecting lumen 120 allows for securement of the elastomeric insert 112 within the counter bore 122. Instead of using a lumen 120, a divot or cut-out may be formed in the back inside wall 117 of the counter bore 122, into which the knob 134 can be placed. In addition to, or instead of, the placement of the knob 134

in the connecting lumen 120, the elastomeric insert 112 may be secured inside the counter bore 122 by use of adhesives or other securement mechanisms.

[0049] The elastomeric joint displayed in Figs. 5A and 5B is configured to mitigate horizontal, axial forces encountered along a curved track in a conveyor system by absorbing and dissipating those forces. When the elastomeric joint 50 is positioned to connect two conveyor carriers as illustrated in Fig. 2, the annular flange 104 of the elastomeric joint 50 is fitted on the vertical post 52 of the front conveyor carrier 58 such that the vertical post 52 extends through the channel 132 in the elastomeric joint 50. The horizontal post 54 on the rear conveyor carrier 60 may be matingly secured in bore 118 in the body portion 114 of the joint housing 102.

[0050] As described above, the forces experienced by the carrier units 14 as they travel through a curve are most significant at the ends of the carrier units 14, at their connection points. When the elastomeric joint 50 experiences axial, horizontal forces in entering and exiting curves on the conveyor track, the elastomeric insert 112 is configured to receive and absorb the forces from the vertical post 52. The rigid vertical post 52 transfers the forces it experiences to the elastomeric insert 112. The physical properties of the elastomeric insert 112 allow it to absorb forces within the confines of the counter bore 122. The elastomeric insert 112 can be deformed (such as compressed or stretched) within the counter bore 122 of the annular flange 104 based upon the forces experienced by the conveyor carrier units, particularly at the vertical post 52. As the elastomeric insert 112 is physically deformed in response to the force received, at least a portion of the force is absorbed by the elastomeric insert 112. The elasticity of the elastomeric insert 112 allows it to regain its original shape after being temporarily deformed in response to the forces received from the vertical post 52.

[0051] Also, because of the gaps 130, 131 present between the inner and outer rings 124, 126 of the elastomeric insert 112, the elastomeric insert 112 is afforded movement in the fore-aft direction, and the elastomeric insert 112 can adjust temporarily according to the force received. Specifically, the inner ring 124 is able to adjust in the fore-aft direction to absorb the forces placed on the inner ring 124 by the vertical post 52. For example, in response to a force received from the vertical post 52 by elastomeric insert 112 in the aft (rear) direction the inner ring 124 will adjust temporarily by moving or bending towards the rear 140 of the elastomeric insert 112, partially filling in the space provided by the rear gap 131. Again, the elasticity of the elastomeric insert 112 allows it to regain its original shape after being temporarily adjusted in response to the forces received from the vertical post 52.

[0052] The fore-aft adjustment of the elastomeric insert 112 within the counter bore 122, along with the force-dissipating material properties of the elastomeric insert 112, allows the

elastomeric joint 50 to successfully absorb and dissipate the horizontal forces experienced by the conveyor train in traveling through a curve. In this manner, the elastomeric joint 50 is capable of keeping the conveyor train taut. That is, by absorbing and dissipating the axial forces experienced by the conveyor units at the joints, the elastomeric joint 50 is also configured to maintain a substantially constant pressure and speed among conveyor units traveling through the curve. Maintaining a constant pressure and speed consequently prevents the dissipation of any forces that are received in the curve through the rest of the conveyor train.

[0053] The bumper 136 positioned in the front gap 130 is particularly suited for dissipating and absorbing forces experienced by the joints when a conveyor train is traveling up and down an inclined conveyor track. The bumper 136 assists in dissipating and absorbing drag forces experienced by the joints by limiting the amount of fore-aft adjustment experienced by the elastomeric insert 112. As a conveyor train travels up an incline, the front 138 of the elastomeric insert 112 will receive a majority of the force applied by the vertical post 52. The inner ring 124 will adjust towards the fore (front side, 138) to compensate for the received force. However, if the inner ring 124 of the elastomeric insert 112 adjusts too much in the fore direction, the over-adjustment may introduce slack into the conveyor train. Thus, to maintain a substantially constant uniform pressure in the elastomeric joints 50, the bumper 136 is included to prevent the inner ring 124 from adjust too much in the fore direction.

[0054] An additional embodiment of an elastomeric insert 212 to be used in an elastomeric joint 50 is displayed in Fig. 6. The elastomeric insert 212 is characterized by a ring having an inside wall 238 and which defines the boundaries of a channel 232 extending therethrough. The elastomeric insert 212 comprises a front side 226 and a rear side 227, as well as lateral sides 230. The thickness of the elastomeric insert 212 between the inside wall 238 and the outer surface is greater in the region defining the lateral sides 230 than it is in the regions comprising the front 226 and rear sides 227. The outside of the elastomeric insert 212 is curved on the lateral sides 230, whereas the outside of the front side 226 and rear side 227 may be flat in portions and may contain a curved section near the middle of the front and rear sides 226, 227. The top 234 and bottom (not displayed in Fig. 6) of the elastomeric insert 212 may be substantially flat. The elastomeric insert 212 may be made of polyurethane, styrene butadiene, natural rubber, urethane, silicone, neoprene, nitrile butadiene, or other such suitable elastomeric material capable of absorbing and dissipating axial forces.

[0055] Fig. 7 is a perspective view of an embodiment of a bearing 242 for use in an elastomeric joint 50. The bearing 242 is comprised of a tubular portion 244 having a top 246, with a channel 248 extending therethrough. The tubular portion 244 is positioned on a top

surface 256 of a base portion 250. The channel 248 formed by the tubular portion 244 extends through the base portion 250 of the bearing 242. The base portion 250 may have two curved sides 258, two flat sides 254, a flat top portion 256 and a flat bottom portion (not shown in Fig. 7). The bearing 242 may be made of sintered or solid metal, plastic, phenolic resin, ceramic materials, etc.

[0056] When this embodiment of the elastomeric joint 50 is assembled, the elastomeric insert 212 and the bearing 242 are placed together in the counter bore 122 of the joint housing 102. To configure the elastomeric insert 212 and bearing 242 in the counter bore 122, the elastomeric insert 212 is placed around the tubular portion 244 of the bearing 242. The tubular portion 244 is configured to extend through the channel 232 of the elastomeric insert 212 and the bottom of the elastomeric insert 212 is configured to rest on the top portion 256 of the base 250 of the bearing 242.

[0057] Figs. 8A, 8B, and 8C provide views of an embodiment of an elastomeric joint 50 in which the elastomeric joint 50 includes a joint housing 202, the elastomeric insert 212, and the bearing 242. Fig. 8A shows the bottom view of an embodiment of an elastomeric joint 50, such that the bottom corresponds to the side of the elastomeric joint 50 facing down towards the track 56 in Fig. 2. Fig. 8B provides a cross-sectional side view of the elastomeric joint 50 displayed in Fig. 8A, taken along line 8B of Fig. 8A. Fig. 8C provides a cross-sectional front view of the elastomeric joint displayed in Fig. 8A, taken along line 8C in Fig. 8A. These cross-sectional views provide additional perspective of the spatial arrangement of the elastomeric insert 212 and the bearing 242 within the counter bore 222 of the annular flange 204.

[0058] In particular, it is seen that the bearing 242 and elastomeric insert 212 are positioned in the counter bore 222 of the annular flange 204 of the joint housing 202; the joint housing 202, the elastomeric insert 212, and the bearing 242 together comprising the elastomeric joint 50. The joint housing 202 is comprised of the body 214 and the annular flange 204. The lateral sides 230 of the elastomeric insert 212 are positioned firmly against the side walls 224 of the counter bore 222. As described above with reference to Figs. 6 and 7, the tubular portion 244 of the bearing 242 is positioned through the channel 232 of the elastomeric insert 212. The curved sides 258 of the bearing 242 are positioned in the direction of the side walls 224 of the counter bore 222 and the flat sides 254 are positioned in the direction of the front and rear interior walls 215, 217 of the counter bore 222. Arranged in this manner, when the elastomeric joint 50 is in a quiescent state and no forces are applied to the elastomeric joint 50, the front 226 of the elastomeric insert 212 is spaced apart from the front inside wall 215 of the counter bore 222 and the rear of the elastomeric insert 212 is spaced apart from the rear

inside wall 217 of the counter bore 222. Thus, viewed from below, as in Fig. 8A, the top 216 of the counter bore 222 may be visible on the front and back sides of the annular flange 204 and the channel 248 in the bearing 242 is centered in the annular flange 204.

[0059] The bottom portion 236 of the elastomeric insert 212 may be bonded or otherwise secured to the top 256 of the base portion 250 of the bearing 242. The top portion 234 of the elastomeric insert 212, likewise, may be bonded or otherwise secured to the top 216 of the counter bore 222 in the annular flange 204 of the joint housing 202. By bonding the elastomeric insert 212 to the bearing 242, and bonding the bearing 242 to the counter bore 222, the elastomeric insert 212 and the bearing 242 are secured in the counter bore 222 and are secured to each other. Yet, the elastomeric insert 212 and the bearing 242 are afforded a measure of fore-aft movement within the counter bore 222 by virtue of the gaps present between the elastomeric insert 212 and the front and rear interior walls 215, 217 of the counter bore 212.

[0060] When the elastomeric joint 50 is used to connect two conveyor carriers, the vertical post 52 on the front conveyor carrier 58 is placed through the channel 248 of the bearing 242 and the horizontal post 54 on the rear conveyor carrier 60 is secured in the bore 218 in the body 214 of the joint housing 202. When the elastomeric joint 50 experiences axial horizontal forces while entering and exiting curves on the conveyor track 56, the forces experienced at the vertical post 52 are transferred to the elastomeric insert 212 and bearing 242. Because of the existence of the gaps present between the elastomeric insert 212 and the front and rear walls 215, 217 of the counter bore 212, the elastomeric insert 212 and the bearing 242 are afforded movement in the fore-aft direction. This fore-aft movement of the elements within the counter bore 222, along with the force-dissipating material properties of the elastomeric insert 212, allows the elastomeric joint 50 to successfully absorb and dissipate the horizontal forces experienced by the conveyor train in traveling through the curve.

[0061] Another advantage of the elastomeric joints disclosed herein includes the dissipation of rotational forces encountered by the conveyor train as one conveyor carrier is banked while a connected conveyor carrier remains flat. For example, when a front conveyor carrier enters a banked portion of the track, the rear conveyor carrier attached behind the front conveyor carrier is still traveling on a flat portion of the track. Such twisting creates rotational forces focused at the joints connecting the conveyor carriers. The rigid, non-elastomeric joints used historically in conveyor systems would propagate these forces through the conveyor train, creating structural stress to the entire conveyor system. The elastomeric joints disclosed herein are configured to absorb and dissipate these rotational forces.

[0062] Another advantage of the elastomeric joints disclosed herein includes the dissipation of vertical forces encountered by the joints connecting conveyor carriers as one conveyor carrier is inclined to begin travel up an inclined portion of the conveyor track while a connected conveyor carrier is flat.

[0063] The embodiments described herein disclose elastomeric joints which dissipate horizontal, rotational, and vertical forces. The efficient dissipation of forces by the elastomeric joints described herein reduces damage to the conveyor joints, the conveyor carriers, and the tracks, and also reduces noise emanating from the conveyor system. Furthermore, because of the reduced damage, the embodiments of the elastomeric joints described herein reduce conveyor system stoppage time, allowing for longer uninterrupted operation of the conveyor system.

[0064] Although illustrated within the context of a conveyor system, embodiments of the disclosure may also be used in other applications. It will be understood by those of skill in the art that numerous and various modifications can be made without departing from the spirit of the development. Therefore, it should be clearly understood that the forms of the development described herein are illustrative only and are not intended to limit the scope of the development.

WHAT IS CLAIMED IS:

1. A joint for connecting carrier units, the joint comprising:

a joint housing comprising:

a head portion on a first side comprising an annular flange, wherein the annular flange is comprised of interior walls that define a hole and a counter bore therethrough;

an elongated body on a second side, wherein the body comprises a bore extending longitudinally therethrough; and

an elastomeric insert positioned inside the counter bore of the annular flange.

2. The elastomeric joint of claim 1, wherein the cross-sectional diameter of the counter bore is greater than the diameter of the hole in the annular flange.

3. The joint of claim 1, wherein the elastomeric insert comprises an outer ring and an inner ring, wherein the inner and outer rings are connected together by supports.

4. The joint of claim 3, further comprising a front and a rear gap in the elastomeric insert, each gap defined by the inner walls of the outer ring and the outer walls of the inner ring.

5. The joint of claim 4, wherein the elastomeric insert further comprises a bumper extending from the inner elastomeric ring and positioned in one of the front or rear gaps.

6. The joint of claim 5, wherein the elastomeric insert further comprises an elastomeric knob extending from an outer surface of the outer ring.

7. The joint of claim 6, wherein the elastomeric insert can be configured in the counter bore to orient the bumper towards a front side of the annular flange and the knob can be configured to fit into a connecting lumen in a rear interior wall of the counter bore.

8. The joint of claim 1 further comprising a bearing, the bearing comprising a base portion and a tubular portion extending from the base portion, wherein the bearing is configured to be positioned inside the counter bore of the annular flange and the elastomeric element configured to fit around the tubular portion of the bearing.

9. The joint of claim 8, wherein a bottom portion of the elastomeric insert is secured to the base portion of the bearing and a top portion of the elastomeric insert is secured to a top of the counter bore of the annular flange.

10. A method of dissipating forces between carrier units, the method comprising:

connecting a first carrier unit to a second carrier unit using an elastomeric joint, wherein the elastomeric joint comprises a joint housing and an elastomeric insert;

receiving a force on the carrier units; and

dissipating the force in the elastomeric joint.

11. The method of claim 10, wherein the step of connecting the first carrier unit to the second carrier unit further comprises:

connecting a first post on the first carrier to an annular flange on the elastomeric joint; and

connecting a second post on the second carrier to a longitudinal bore in the elastomeric joint.

12. The method of claim 11, wherein the step of dissipating the force in the elastomeric joint further comprises receiving the force experienced by the first carrier at the first post at the elastomeric insert.

13. The method of claim 12, wherein the elastomeric insert is configured to dissipate the force by elastically adjusting within the joint housing and by absorbing the force.

14. The method of claim 12, further comprising connecting a plurality of carrier units using a plurality of elastomeric joints, wherein the plurality of carrier units comprise a conveyor train.

15. The method of claim 14, wherein the step of receiving a force is a result of the plurality of carrier units traveling in a curve.

16. The method of claim 14, wherein the step of receiving a force is a result of the plurality of carrier units traveling on an incline.

17. A joint for dissipating forces in carrier units, the joint comprising:

means for connecting conveyor units;

means for receiving a force on the conveyor units; and

means for dissipating the force within the connecting means.

18. The elastomeric joint of claim 17, wherein the connecting means comprises a elastomeric joint.

19. The elastomeric joint of claim 17, wherein the dissipating means comprises an elastomeric insert.

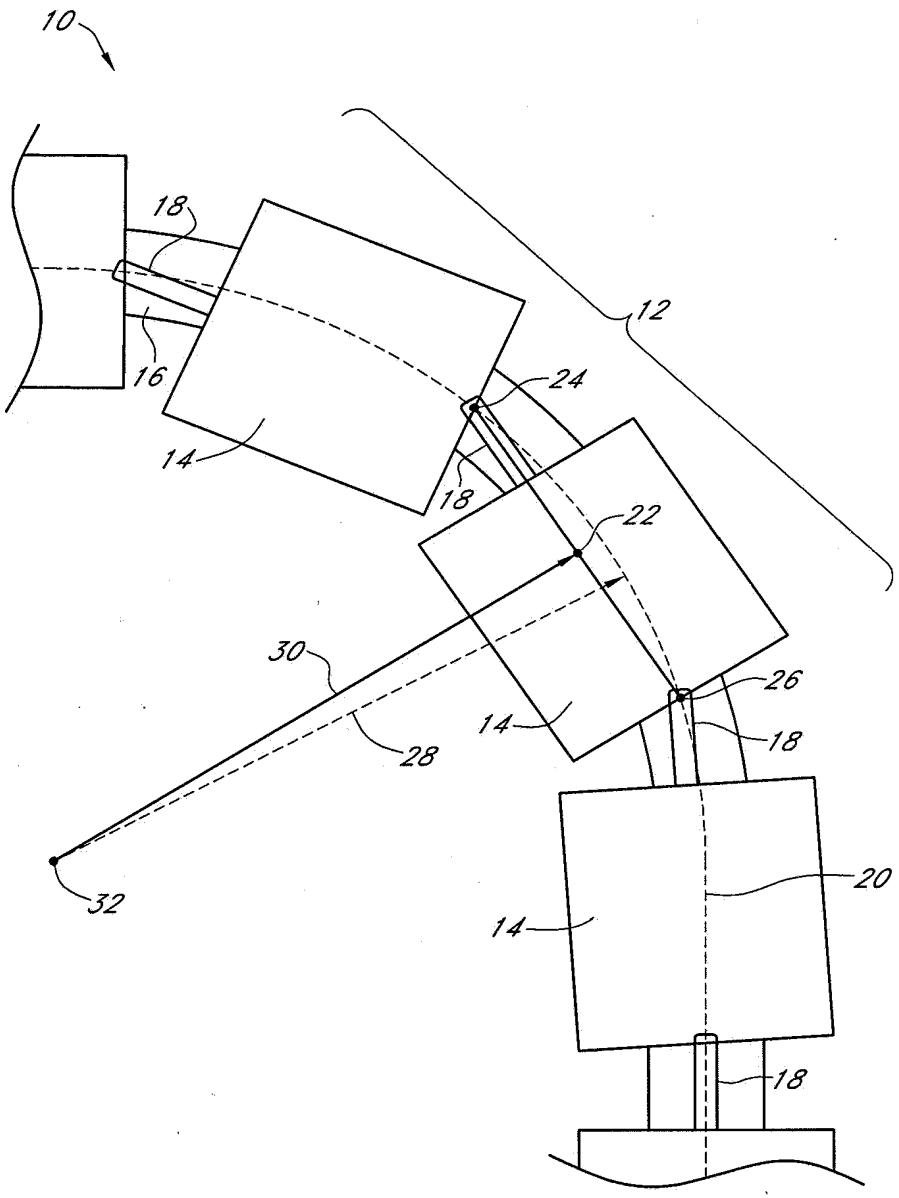


FIG. 1

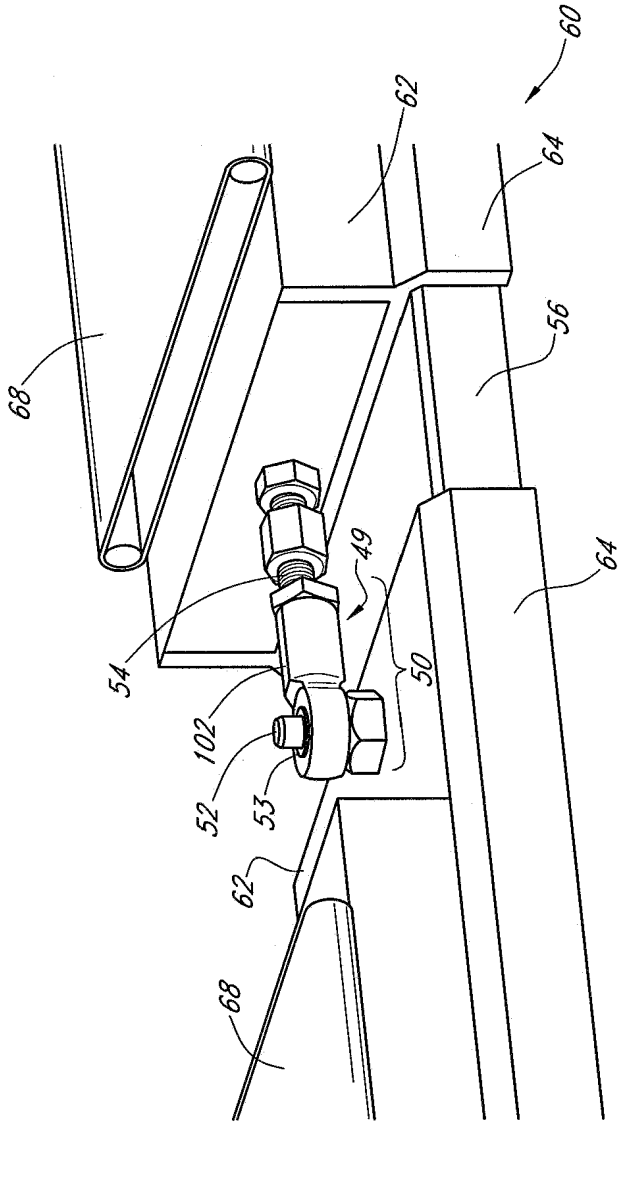


FIG. 2

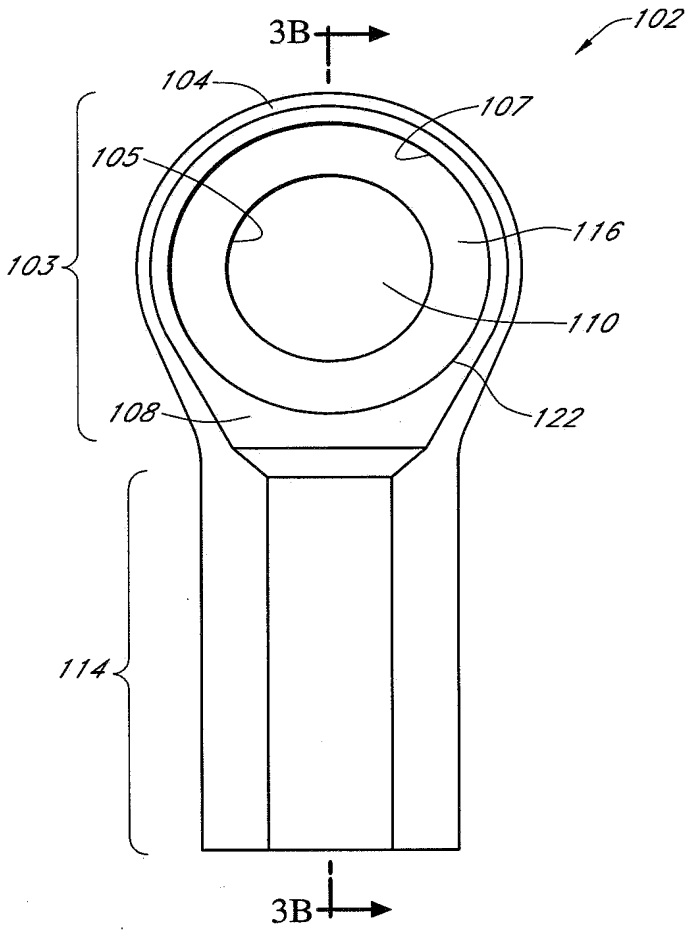


FIG. 3A

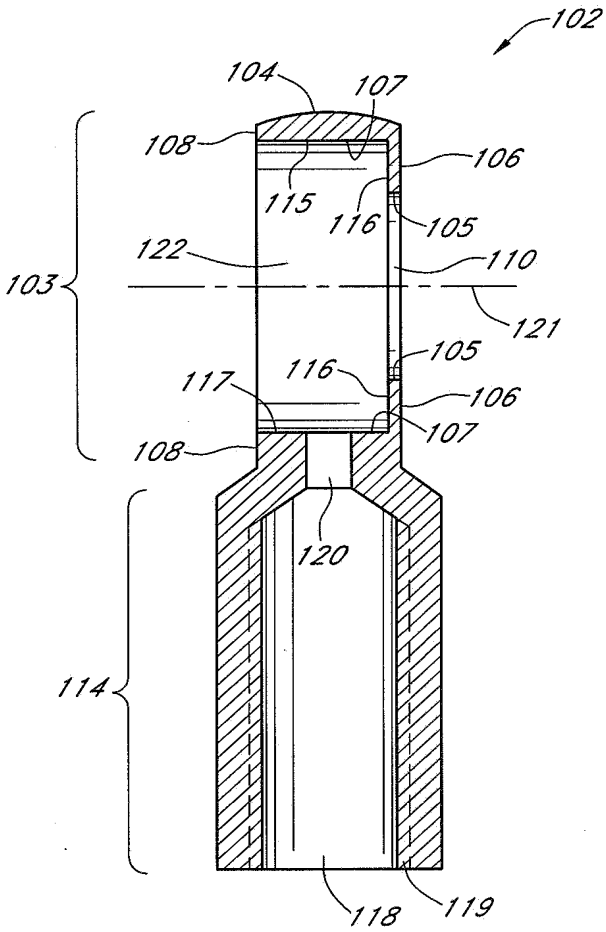

102
104
107
108
106
115
116
105
122
110
103
121
105
116
117
106
107
108
120
114
118
119

FIG. 3B

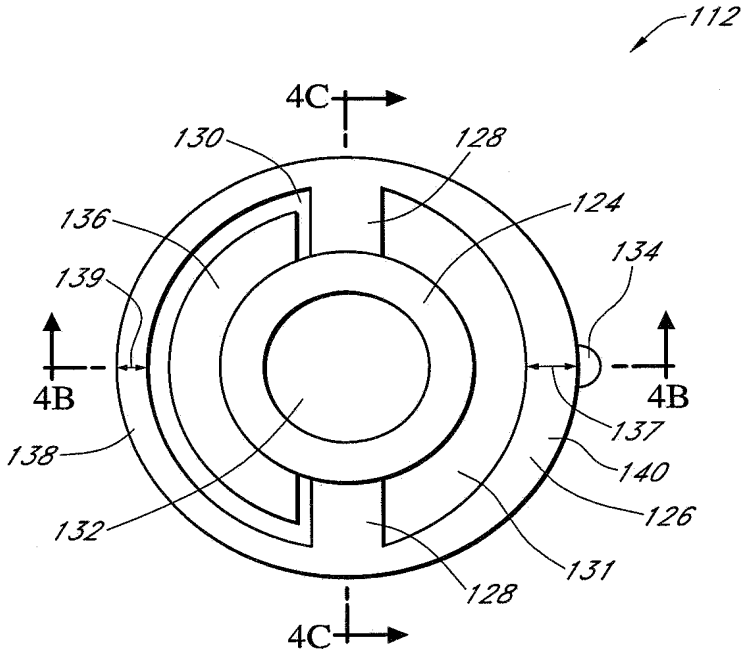

112
4C
130
128
124
136
134
139
4B
4B
137
138
140
126
132
131
128
4C

FIG. 4A

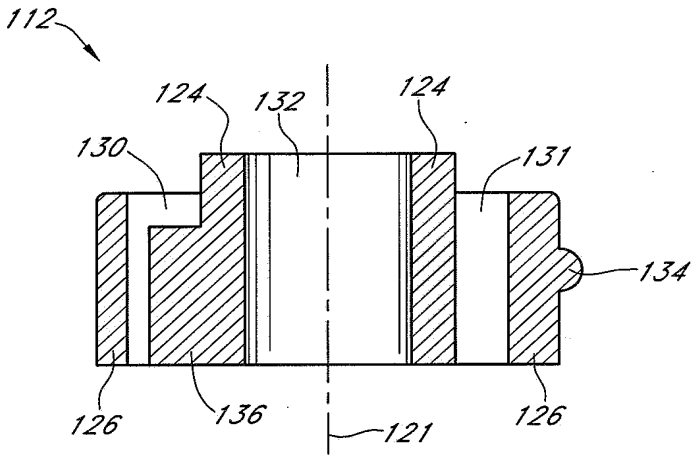

112
124
132
124
130
131
134
126
136
121
126

FIG. 4B

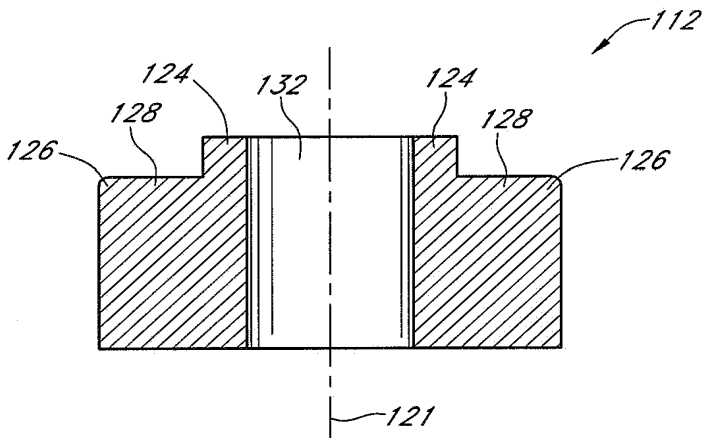

112
124
132
124
128
128
126
126
121

FIG. 4C

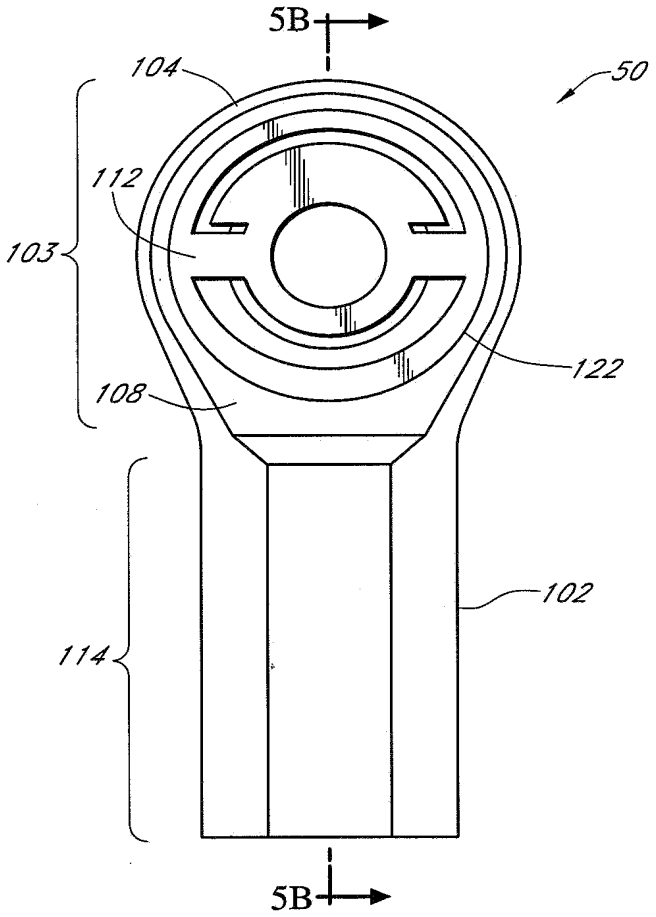


FIG. 5A

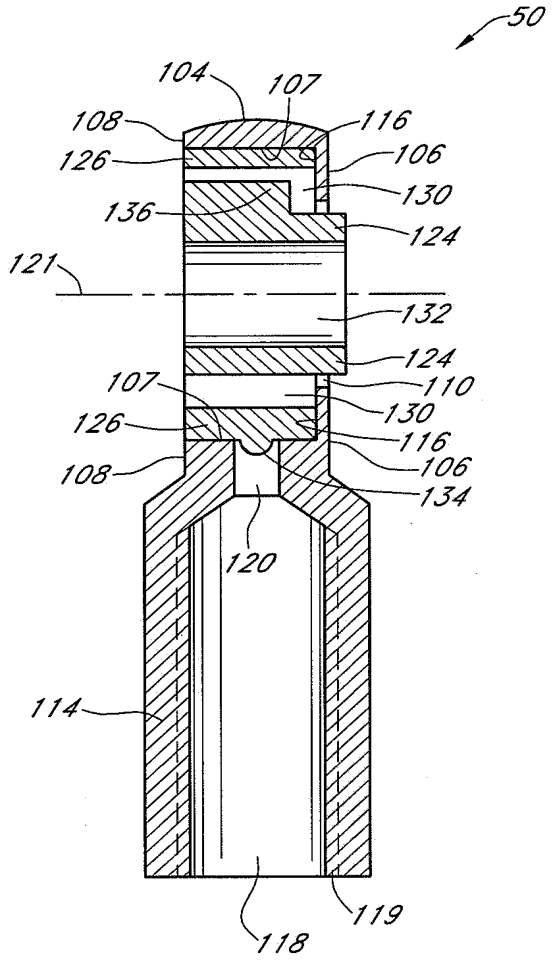

50
104
107
108
116
126
106
130
136
124
121
132
124
107
110
130
126
116
106
108
134
120
114
119
118

FIG. 5B

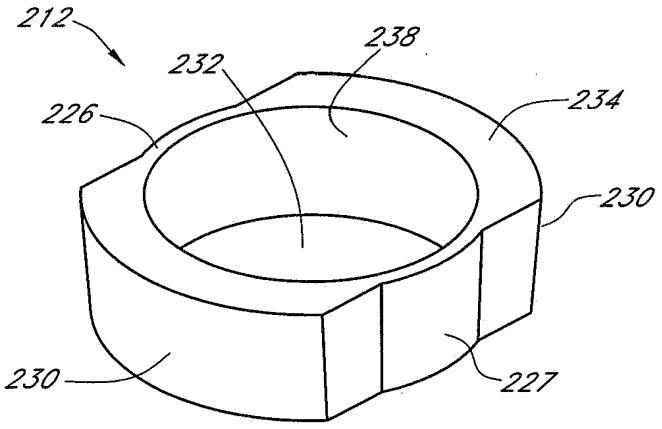

212
238
232
234
226
230
230
227

FIG. 6

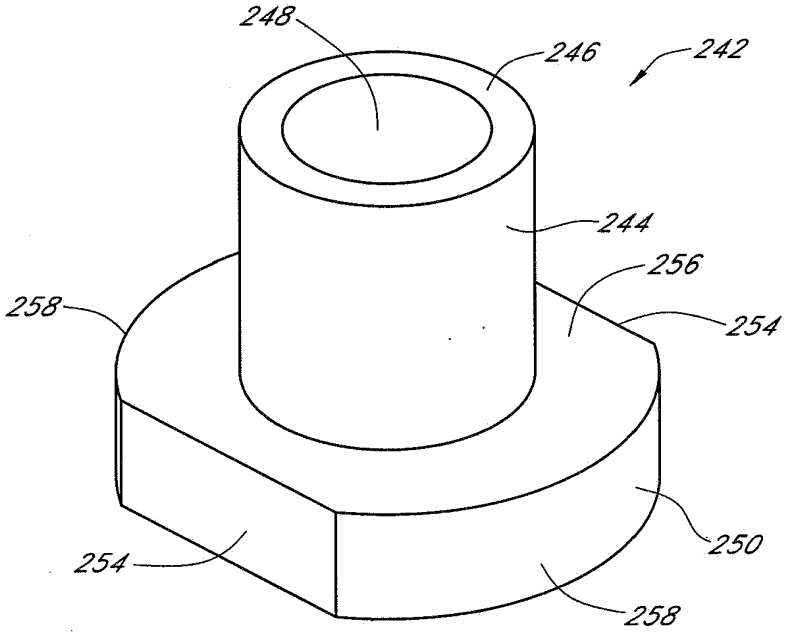

248
246
242
244
256
258
254
250
254
258

FIG. 7

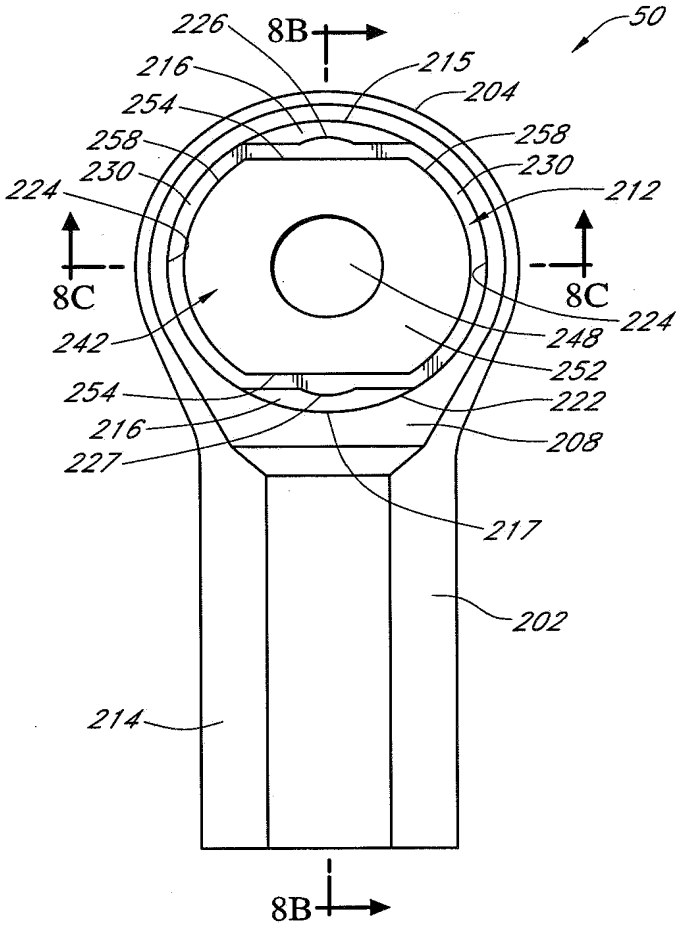

50
226
8B
215
216
204
254
258
258
230
230
224
212
8C
8C
224
242
248
252
254
222
216
208
227
217
202
214
8B

FIG. 8A

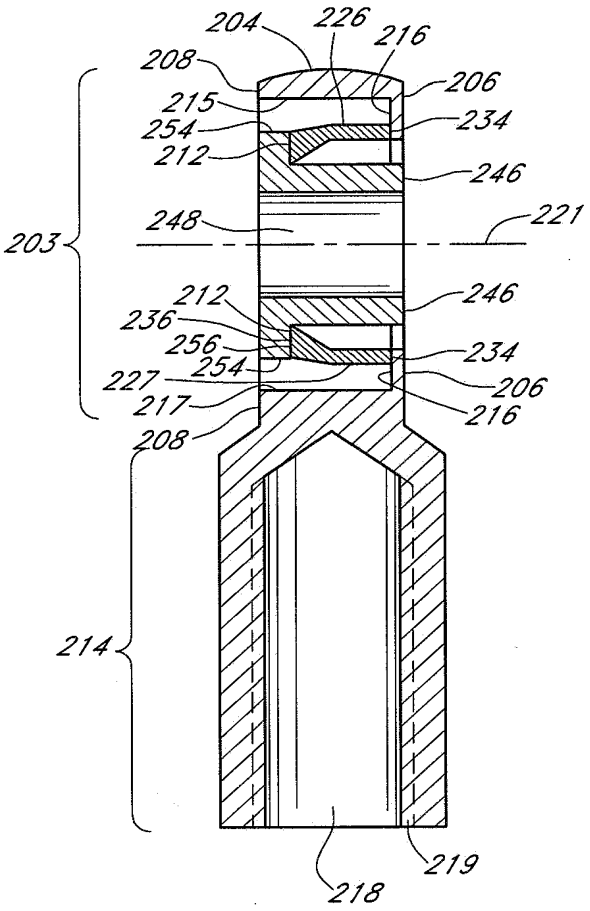

204
226
216
208
206
215
234
254
212
246
248
221
203
212
246
236
256
234
254
227
206
217
216
208
214
218
219

FIG. 8B

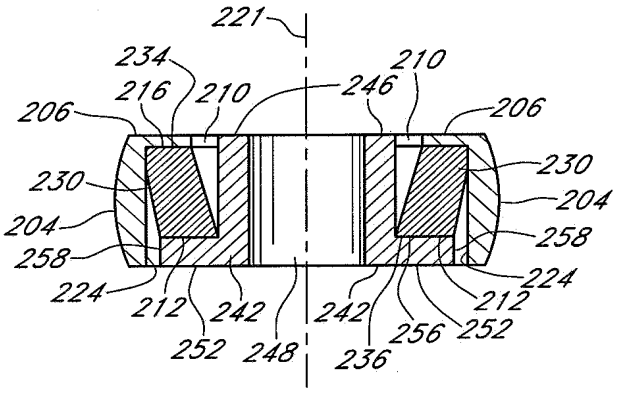

221
234
210
216
210
246
206
206
230
230
204
204
258
258
224
212
242
242
212
252
248
236
256
252

FIG. 8C

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2012/040926

A. CLASSIFICATION OF SUBJECT MATTER
INV. B65G17/22 B65G47/96
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)
B65G

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 0 671 349 A1 (PROMECH SORTING SYSTEMS) 13 September 1995 (1995-09-13)	10,17-19
A	column 4, line 32 - column 6, line 13; figures	1,12-16

A	US 5 680 924 A (GRAPHA-HOLDING) 28 October 1997 (1997-10-28) the whole document	1

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
5 September 2012	27/09/2012

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	Jagusiak, Antony

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

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2012/040926

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 671349 A1	13-09-1995	AT 157622 T	15-09-1997
		AU 678675 B2	05-06-1997
		AU 1369595 A	21-09-1995
		CA 2144301 A1	12-09-1995
		DE 69500629 D1	09-10-1997
		DE 69500629 T2	08-01-1998
		EP 0671349 A1	13-09-1995
		JP 7304512 A	21-11-1995
		NL 9400388 A	02-10-1995
US 5680924 A	28-10-1997	AT 168966 T	15-08-1998
		CH 687616 A5	15-01-1997
		DE 59502967 D1	03-09-1998
		DK 711244 T3	03-05-1999
		EP 0711244 A1	15-05-1996
		JP H09501386 A	10-02-1997
		US 5680924 A	28-10-1997
		WO 9533669 A1	14-12-1995

Espacenet

Bibliographic data: WO2014153757 (A1) — 2014-10-02

METHOD, SYSTEM, AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION OF PREDETERMINED AREA IN FETAL GENOME

Inventor(s): YIN XUYANG [CN]; JIANG HUI [CN]; CHEN SHENGPEI [CN]; GONG CHUN [CN]; CHEN FANG [CN]; ZHANG CHUNLEI [CN]; PAN XIAOYU [CN] ± (YIN, XUYANG, ; JIANG, HUI, ; CHEN, SHENGPEI, ; GONG, CHUN, ; CHEN, FANG, ; ZHANG, CHUNLEI, ; PAN, XIAOYU)

Applicant(s): BGI SHENZHEN [CN] ± (BGI SHENZHEN)

Classification: **- international:** ***C12Q1/68***
- cooperative: **C12Q1/6876 (EP); G16B20/00 (EP); G16B30/00 (EP);** C12Q2600/156 (EP)

Application number: WO2013CN73375 20130328 Global Dossier

Priority number(s): WO2013CN73375 20130328

Also published as: CN105051208 (A) CN105051208 (B) HK1213945 (A1)

Abstract of WO2014153757 (A1)

A method, a system, and a computer readable medium for determining base information of a predetermined area in a fetal genome are provided. The method for determining base information of a predetermined area in a fetal genome comprises the following steps: obtaining a sequencing result of a DNA sample of a genome of an embryo cell, and a sequencing result of a genome sample of an individual related to embryo heredity; constructing a heredity sketch of the embryo based on the sequencing result of the DNA sample of the genome of the embryo cell, to determine an initial genotype of the embryo; determining haplotypes of parents of the embryo based on the sequencing result of the genome sample of the individual related to the embryo heredity; and using the initial genotype of the embryo as an observation sequence according to the hidden markov model, and determining the base information of the predetermined area in the fetal genome based on the haplotypes of the parents of the embryo.

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2014 年 10 月 2 日 (02.10.2014)

WIPO | PCT

(10) 国际公布号
WO 2014/153757 A1

(51) 国际专利分类号:
C12Q 1/68 (2006.01)

(21) 国际申请号: PCT/CN2013/073375

(22) 国际申请日: 2013 年 3 月 28 日 (28.03.2013)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: **深圳华大基因研究院 (BGI SHENZHEN)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(72) 发明人: **殷旭阳 (YIN, Xuyang)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **蒋慧 (JIANG, Hui)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **陈盛培 (CHEN, Shengpei)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **龚淳 (GONG, Chun)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **陈芳 (CHEN, Fang)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **张春雷 (ZHANG, Chunlei)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **潘小瑜 (PAN, Xiaoyu)**; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: **北京清亦华知识产权代理事务所（普通合伙） (TSINGYIHUA INTELLECTUAL PROPERTY LLC)**; 中国北京市海淀区清华园清华大学照澜院商业楼 301 室, Beijing 100084 (CN)。

(81) **指定国** (除另有指明，要求每一种可提供的国家保护): 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, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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) **指定国** (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

(54) **Title**: METHOD, SYSTEM, AND COMPUTER READABLE MEDIUM FOR DETERMINING BASE INFORMATION OF PREDETERMINED AREA IN FETAL GENOME

(54) **发明名称** : 确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质

(57) **Abstract**: A method, a system, and a computer readable medium for determining base information of a predetermined area in a fetal genome are provided. The method for determining base information of a predetermined area in a fetal genome comprises the following steps: obtaining a sequencing result of a DNA sample of a genome of an embryo cell, and a sequencing result of a genome sample of an individual related to embryo heredity; constructing a heredity sketch of the embryo based on the sequencing result of the DNA sample of the genome of the embryo cell, to determine an initial genotype of the embryo; determining haplotypes of parents of the embryo based on the sequencing result of the genome sample of the individual related to the embryo heredity; and using the initial genotype of the embryo as an observation sequence according to the hidden markov model, and determining the base information of the predetermined area in the fetal genome based on the haplotypes of the parents of the embryo.

(57) **摘要**: 提供了确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质。其中，确定胚胎基因组中预定区域碱基信息的方法，包括下列步骤：获得胚胎细胞基因组 DNA 样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果；基于胚胎细胞基因组 DNA 样本的测序结果，构建该胚胎的遗传草图，以便确定胚胎初始基因型；基于胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以该胚胎初始基因型作为观察序列，基于该胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

WO 2014/153757 A1

确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质

优先权信息

无

技术领域

本发明涉及确定胚胎基因组中预定区域碱基信息的方法、系统和计算机可读介质。

背景技术

遗传性疾病是由于遗传物质发生改变而造成的疾病，具有先天性、家族性、终身性和遗传性的特点。遗传性疾病可分为 3 个大类：单基因遗传病、多基因遗传病及染色体异常。其中单基因病多由于单个致病基因的显性或隐性遗传所致基因功能异常；而多基因遗传病则是由多个基因变化影响所致的疾病，会在一定程度上受到外界环境因素的影响；染色体异常包括数目异常和结构异常，最为多见的是由于第 21 号染色体三体所致的唐氏综合症，患儿表现为先天愚型和肢体形状异常等其他先天性特征。由于目前对遗传性疾病尚无有效的治疗方式，只能针对性地进行支持治疗或者药物缓解，费用昂贵，给社会和家庭带来沉重经济和精神负担。因此，在孩子出生前就对孩子的患病状态进行检测，做好预防工作，以达到优生优育的目的，是十分必要的。例如在体外生殖过程中，在胚胎植入子宫之前，对胚胎的患病状态进行检测，将具有重大意义。对配子或植入到子宫腔之前的胚胎进行遗传学分析，将排除致病因素的正常胚胎植入子宫，从而防止患有遗传病的孩子的妊娠。对具有高风险妊娠遗传缺陷胚胎的夫妇实施植入前胚胎的患病状态的检测，无疑能有效降低患病孩子的出生，并且避免选择性终止妊娠给孕妇及家庭带来的诸多危害。

然而，目前的相关检测手段仍有待改进。

发明内容

本发明是基于发明人的下列发现而完成的：

发明人发现，植入前胚胎中可以活检取样进行检测的细胞数有限。例如培养至第三天的卵裂球仅有 4-8 个细胞，可以进行活检取样的仅 1-2 个细胞。即使在培养至第五天的囊胚，进行外滋养层细胞活检取样也只能取到 3-10 个细胞。由于单个或少数胚胎细胞的 DNA 量有限，难以直接进行全面的遗传变异检测，必需经过全基因组扩增(Whole Genome Amplification, WGA)，达到足够 DNA 量，才能对遗传变异做全面分析。而全基因组扩增常存在偏好性，会产生一定的杂合缺失或等位基因脱扣(Allele Drop Out)等现象，这些都为胚胎细胞的遗传病相关的遗传变异检测带来风险。

因此，很有必要开发能够对扩增引起的等位基因脱扣等错误进行定位和矫正的方法，以提高遗传变异检测的准确性。

本发明旨在至少解决现有技术中存在的技术问题之一。通过本发明方法推算胚胎单

倍体型，可以对发生等位基因脱扣的位点进行定位，并能够进行矫正。而胚胎单倍体型分析方法可以基于家系成员如父母和先证者的遗传信息，或者基于来自同一父母的其他胚胎细胞的遗传信息进行推算。

在本发明的一个方面，本发明提出了一种确定胚胎基因组中预定区域碱基信息的方法。该方法结合系谱推断和统计算法来进行胚胎单体型的确定，通过家系中相关个体的基因型，即追溯染色体片段的传递，结合统计算法来推断单体型状态。根据本发明的实施例，该方法包括下列步骤：获得胚胎细胞基因组 DNA 样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果；基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

需要说明的是，子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。将胚胎的单倍型作为隐含状态（hidden states），将胚胎单细胞全基因组扩增后的测序数据作为观察序列（observations），利用贝叶斯算法借助先验数据推算出状态转移概率（transition probabilities），构建观察序列概率分布（observation symbol probabilities）和初始状态概率分布（initial state distribution），然后通过惠特比算法（Viterbi algorithm）可以推断出最可能的胚胎单体型组合。因而，根据本发明的实施例，借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。对于在胚胎细胞全基因组扩增 DNA 测序结果中覆盖程度较低，存在杂合缺失或等位基因脱扣(Allele Drop Out)的位点或 DNA 序列，可以通过本方法准确地推断得到。因此通过本方法，可以对杂合缺失或等位基因脱扣(Allele Drop Out)的位点或序列进行校正，使检测结果更加准确可靠。

在本发明的又一方面，本发明提出了一种用于确定胚胎基因组中预定区域碱基信息的系统。根据本发明的实施例，该系统包括：文库构建装置，所述文库构建装置适于针对胚胎基因组 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库；测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得胚胎以及胚胎遗传相关个体的测序结果；以及分析装置，所述分析装置与所述测序装置相连，并且适于：基于胚胎基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。根据本发明的实施例，还可以进一步包括胚胎活检装置，所述胚胎活检装置适于针对体外培养的胚胎进行活检，对胚胎细胞进行取样。利用该系统，能够有效地实施前面所述的确定胚胎基因组中预定区域碱基信息的方法，可以借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序

列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测，从而可以有效地对胚胎基因组的遗传信息进行确定。

在本发明的另一方面，本发明还提出了一种计算机可读介质。根据本发明的实施例，基于胚胎细胞基因组DNA样本的测序结果，构建胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及以根据隐马尔可夫模型，所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。利用本发明的计算机可读介质，能够有效地被处理器执行其存储的指令，以便借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），基于胚胎细胞的测序结果，参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

本发明的上述和/或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

图1为根据本发明一个实施例的利用隐马尔可夫模型进行分析的流程示意图；以及

图2为根据本发明的一个实施例的用于确定胚胎基因组中预定区域核酸序列的系统的结构示意图。

发明详细描述

下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。

需要说明的是，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。进一步地，在本发明的描述中，除非另有说明，“多个”的含义是两个或两个以上。

确定胚胎基因组中预定区域碱基信息的方法

在本发明的第一方面，本发明提出了一种确定胚胎基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：

首先，获得胚胎细胞基因组DNA样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果。其中，在本文中所使用的术语“胚胎遗传相关个体”指的是在遗传意义上，与胚胎之间具有亲缘关系的个体，例如根据本发明的实施例，可以采用的“胚胎遗传相关个体”为胚胎的亲代例如父母、祖父母、外祖父母以及胚胎父母的其他子女。这里所说的

胚胎父母的其他子女应做广义理解，既可以是已经出生的子女，也可以是尚未出生的子女（受精卵或胚胎），也可以是死亡的胚胎，也可以是体外培养的胚胎或受精卵，只要与待检测胚胎共有相同的父母即可。

根据本发明的实施例，胚胎细胞基因组 DNA 样本的来源并不受特别限制。根据本发明的实施例，具体地，可以针对胚胎进行活检取样，获得胚胎细胞，并对胚胎细胞样本进行全基因组扩增(WGA)，以便获得胚胎细胞基因组 DNA。在本文中所采用的术语“胚胎活检”是指利用显微技术，从胚胎中分离部分胚胎细胞，或从受精卵/配子中分离部分细胞的技术。其中，根据本发明的实施例，该胚胎细胞可来自于卵裂球、囊胚滋养层、受精卵和配子中的任一种，可以是单个细胞，也可以是 2-10 个细胞。此外，也可以采用任何含有胚胎核酸的孕妇样本进行全基因组扩增，以便获得胚胎基因组 DNA 样本。由此，可以在不影响胚胎正常发育的前提下，对胚胎的基因组进行有效监测。其中，“全基因组扩增(WGA)”主要包括多重链置换扩增(Multiple Displacement Amplification, MDA)和基于 PCR 的 WGA 技术，可以采用自主开发的 WGA 扩增流程，也可以采用商业化的试剂盒如 Qiagen 公司的 REPLI-g 系列的试剂盒、Sigma-Aldrich 公司的 GenomePlex WGA 试剂盒(WGA4)、Rubicon Genomics 公司的 PicoPlex WGA 试剂盒、GE-Healthcare 公司的 illustra Genomiphi WGA 试剂盒等。在扩增 DNA 样本之后，针对胚胎细胞的扩增 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库。

关于针对核酸样本，构建测序文库的方法和流程，本领域技术人员可以根据不同的测序技术进行适当选择，关于流程的细节，可以参见测序仪器的厂商例如 Illumina 公司所提供的规程，例如参见 Illumina 公司 Multiplexing Sample Preparation Guide （Part#1005361; Feb 2010）或 Paired-End Sample Prep Guide（Part#1005063; Feb 2010），通过参照将其并入本文。根据本发明的实施例，从生物样本提取核酸样本的方法和设备，也不受特别限制，可以采用商品化的核酸提取试剂盒进行。

在构建测序文库后，将测序文库应用于测序仪器，对测序文库进行测序，并获得相应的测序结果，该测序结果是由多个测序数据构成的。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，能够提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46）包括但不限于 Illumina-Solexa（GA™,HiSeq2000™, Miseq 等）、ABI-SOLiD、Life Technologies-Ion Torrent/Proton 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing），Pacific Biosciences 公司单分子实时测序（single molecule real-time (SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245）。随着测序

技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。

根据本发明的实施例，在得到测序结果之后，可以基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型。基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型。根据本发明的实施例，通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图。通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。根据本发明的实施例，可以采用已知的人类参考基因组作为参考序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI 36.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

最后，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息。

在本文中所使用的术语“预定区域”应作广义理解，是指任何包含可能发生预定事件位点的核酸分子的区域。对于 SNP 分析而言，可以是指包含 SNP 位点的区域。对于分析染色体非整倍性，则预定区域指的是所要分析的染色体的全长或者部分，即选择所有来自该染色体的测序数据。从测序结果中选择来自相应区域的测序数据的方法可以不受特别限制。根据本发明的实施例，可以通过将所得到的所有测序数据与已知的核酸参考序列进行比对，从而得到来自于预定区域的测序数据。另外，根据本发明的实施例，预定区域也可以是基因组上不连续的多个分散点。根据本发明的实施例，可以使用的参考序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。

根据本发明的实施例，可以通过基于所述胚胎细胞的测序结果，结合胚胎遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，可以借助隐马尔可夫模型，利用惠特比算法（Viterbi algorithm），确定胚胎基因组中特定区域的碱基信息。由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

下面参考图 1，对借助隐马尔可夫模型利用惠特比算法进行分析的原理进行详细描述：

如前所述，在本文中所使用的术语“胚胎遗传相关个体”指的是在遗传意义上，与胚胎之间具有亲缘关系的个体，例如根据本发明的实施例，可以采用的“胚胎遗传相关个体”为胚胎的亲代例如父母、祖父母、外祖父母以及胚胎父母的其他子女。这里所说的胚胎父母的其他子女应做广义理解，既可以是已经出生的子女，也可以是尚未出生的子女（胚胎或受精卵），也可以是死亡的胚胎，也可以是体外培养的胚胎或受精卵，只要与待检测胚胎共有相同的父母即可。

由此，子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。将胚胎的单倍型作为隐含状态（hidden states），将胚胎

单细胞全基因组扩增后的测序数据作为观察序列（observations），借助先验数据推算出状态转移概率（transition probabilities），构建确定观察序列概率分布（observation symbol probabilities）和初始状态概率分布（initial state distribution），然后通过惠特比算法（Viterbi algorithm）可以推断出最可能的胚胎单体型组合。因而，根据本发明的实施例，借助隐马尔可夫模型，例如可以通过利用惠特比算法（Viterbi algorithm），参考胚胎遗传相关个体的遗传信息，可以确定胚胎基因组中特定区域的核酸序列，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

详细分析步骤如下:

标记:

I. 需要检测的位点数为 N。

II. 观测序列;

III. 隐藏状态集为: $S=\{0,1\}$，定义了父母哪条染色体遗传给了子代。0 代表遗传给了先证者的那条染色体，1 代表没有遗传给先证者的那条染色体。

IV. 观测状态集为: $V=\{0,1\}$，1 代表遗传给先证者的染色体与胚胎遗传草图基因型一致，0 代表不一致。

第一步，构建初始状态概率分布向量，单倍体重组转移矩阵以及观察序列概率矩阵

I. 初始状态概率分布记为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

在没有参考数据的情况下，可以假设每种隐藏状态出现的可能性相等。

II. 记单倍体重组转移矩阵记为 $A=\{a_{ij}\},(i,j\in S)$，即隐藏状态转移的概率。

$$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases},$$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，根据先验数据 Nr 可取 20-40 之间的自然数，用来计算隐藏状态的转移概率，即计算每个碱基组合的单体型发生重组的概率。

III. 记观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，这里的概率分布 $b_i(k)$ 用子代一定是纯合（Must-hom）或一定是杂合(Must-het)的位点来估算。

$b_i(0)=\dfrac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$, $b_i(1)=1-b_i(0)$。

第二步，构建局部概率矩阵和逆向指针

定义 局部概率 : $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$。

定义 逆向指针：$\Psi_i(j)=\arg\max \delta_{t-1}(i)\cdot a_{ij} \quad t\in\{1...N\}$。

第三步，确定最终状态状态，并回溯最优路径

确定最终状态，$q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$。

回溯最优路径，即最可能胚胎单基因型 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

第四步，输出结果

即，根据本发明的实施例，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

根据具体的实施例，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 为自然数，取值范围 20-40，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$, $b_i(1)=1-b_i(0)$,

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K) \quad t\in\{1...N\}$,

逆向指针为 $\Psi_i(j)=\arg\max \delta_{t-1}(i)\cdot a_{ij} \quad t\in\{1...N\}$,

依惠特比算法经递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$,

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

其中，这里所使用的术语“局部概率$\delta_i(q_i)$”和“逆向指针$\Psi_i(q_i)$”都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE， Vol.77，No.2，1989 年 2 月，通过参照将其全文并入本文。

由此，能够有效地对胚胎基因组的序列进行分析。相比其他已有的植入前检测技术方法，本方法有以下技术优势，主要体现在准确性和可获得的遗传信息量上:

1) 根据本发明实施例，针对胚胎的父源性位点和母源性位点，均可以很好地检测，检测准确率可高达 95%以上，且可以检测多种变异类型，扩大了疾病检测的范围。

2) 根据本发明实施例，对于一些在胚胎单细胞或微量细胞全基因组扩增 DNA 测序结果中覆盖程度较低，存在杂合缺失或等位基因脱扣(Allele Drop Out)的位点或 DNA 序列，可以通过本方法准确地推断得到。因此通过本方法，可以对杂合缺失或等位基因脱扣(Allele Drop Out)的位点或序列进行校正，使检测结果更加准确可靠。

3) 根据本发明实施例，可进行遗传疾病作图，对于一些连锁相关疾病，可通过其他位点的信息直接推断出来，一次可获得的信息量大，对临床检测更加具有指导意义。

另外，根据本发明实施例，本发明的确定胚胎基因组中预定区域碱基信息的方法，不仅限于 SNP 或者 STR 等某一种遗传多态性位点，对所有的遗传多态性位点均可适用，且可以多种位点同时使用，以便互相验证。

用于确定胚胎基因组中预定区域碱基信息的系统

在本发明的又一方面，本发明提出了一种用于确定胚胎基因组中预定区域核酸序列的系统。根据本发明的实施例，参考图 2，该系统 1000 可以包括: 文库构建装置 100、测序装置 200 以及分析装置 400。

根据本发明的实施例，文库构建装置 100 适于针对胚胎基因组 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库。根据本发明的实施例，测序装置 200 与文库构建装置 100 相连，并且适于对所述测序文库进行测序，以便获得胚胎以及胚胎遗传相关个体的测序结果。根据本发明的实施例，还可以进一步包括 DNA 样本分离装置和 DNA 扩增装置（图中未示出），该 DNA 样本分离装置适于针对胚胎进行活检取样，获得胚胎细胞。胚胎细胞可来自于卵裂球、囊胚滋养层、受精卵和配子中的任一种，可以是单个细胞，也可以是微量细胞例如 2-10 个细胞，也可以采用任何含有胚胎核酸的孕妇样本。该 DNA 扩增装置适于针对活检取样获得的胚胎细胞进行全基因组扩增，以获得足够量 DNA 用于后续检测分析。由此，可以在不影响胚胎正常发育的前提下，对胚胎的基因组进行有效监测。关于全基因组扩增的方法和流程，主要包括多重链置换扩增(Multiple Displacement Amplification, MDA)和基于 PCR 的 WGA 技术，可以采用自主开发的 WGA 扩增流程，也可以采用商业化的试剂盒如 Qiagen 公司的 REPLI-g 系列的试剂盒、Sigma-Aldrich 公司的 GenomePlex WGA 试剂盒(WGA4)、Rubicon Genomics 公司的 PicoPlex WGA 试剂盒、GE-Healthcare 公司的 illustra Genomiphi WGA 试剂盒等等。

关于针对核酸样本，构建测序文库的方法和流程，本领域技术人员可以根据不同的测序技术进行适当选择，关于流程的细节，可以参见测序仪器的厂商例如 Illumina 公司所提供的规程，例如参见 Illumina 公司 Multiplexing Sample Preparation Guide （Part#1005361; Feb 2010）或 Paired-End SamplePrep Guide（Part#1005063; Feb 2010），通过参照将其并入本文。根据本发明的实施例，从生物样本提取核酸样本的方法和设备，也不受特别限制，可以采用商品化的核酸提取试剂盒进行。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，能够提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台（Metzker ML. Sequencing technologies-the next generation. Nat Rev Genet. 2010 Jan;11(1):31-46）包括但不限于 Illumina-Solexa（GA™、HiSeq™、Miseq 等）、ABI-SOLiD、Life Technologies-Ion Torrent/Proton 和 Roche-454（焦磷酸测序）测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术（True Single Molecule DNA sequencing），Pacific Biosciences 公司单分子实时测序（single molecule real-time (SMRT™)），以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等（Rusk, Nicole (2009-04-01). Cheap Third-Generation Sequencing. Nature Methods 6 (4): 244-245）。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。

根据本发明的实施例，还可以包括比对装置 300。根据本发明的实施例，比对装置 300 与测序装置 200 相连，并且适于将所得到的测序结果与参考序列进行比对，以便通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图；通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。根据本发明的实施例，可以使用的参考序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参考序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI 36.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

根据本发明的实施例，分析装置 400 适于：基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

根据本发明的实施例，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息

进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

根据具体的实施例，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr, Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 可取 20-40 之间的自然数，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，

$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$, $b_i(1)=1-b_i(0)$,

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$,

逆向指针为 $\Psi_t(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$,

递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$,

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

这里所使用的术语局部概率 $\delta_i(q_i)$和逆向指针 $\Psi_i(q_i)$都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE， Vol.77，No.2，1989 年 2 月，通过参照将其全文并入本文。

关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胚胎基因组中预定区域核酸序列的系统。不再赘述。

由此，利用该系统，能够有效地实施前面所述的确定胚胎基因组中预定区域核酸序列的方法，可以借助隐马尔可夫模型，通过例如惠特比算法（Viterbi algorithm）解码，确定胚胎基因组中特定区域的碱基信息，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

在本文中所述的术语“相连”应作广义理解，既可以是直接相连，也可以是间接相连，只要能够实现上述功能上的衔接即可。

需要说明的是，本领域技术人员能够理解，在前面所描述的确定胚胎基因组中预定区域核酸序列的方法的特征和优点也适合于确定胚胎基因组中预定区域核酸序列的系统，为描述方便，不再详述。

计算机可读介质

在本发明的又一方面，本发明提出了一种计算机可读介质。根据本发明的实施例，所述计算机可读介质上存储有指令，所述指令适于被处理器执行以便:

基于胚胎细胞基因组 DNA 样本的测序结果，构建胚胎的遗传草图，以便确定胚胎初始基因型;

基于该胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以该胚胎初始基因型作为观察序列，基于该胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

由此，利用该计算机可读介质，能够有效地实施前面所述的方法，从而可以通过例如惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胚胎基因组中特定区域的碱基信息，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

根据本发明的实施例，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

根据具体的实施例，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$，

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 可取 20-40 之间的自然数，观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，

$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$，$b_i(1)=1-b_i(0)$，

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K) \quad t\in\{1...N\}$

逆向指针为 $\Psi_i(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij} \quad t\in\{1...N\}$，

经递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$，

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

这里所使用的术语局部概率$\delta_i(q_i)$和逆向指针$\Psi_i(q_i)$都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 Lawrence R. Rabiner, PROCEEDINGS OF THE IEEE， Vol.77，No.2，1989 年 2 月，通过参照将其全文并入本文。

关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胚胎基因组中预定区域核酸序列的系统。不再赘述。

由此，利用该系统，能够有效地实施前面所述的确定胚胎基因组中预定区域核酸序列的方法，可以通过例如惠特比算法（Viterbi algorithm），借助隐马尔可夫模型，确定胚胎基因组中特定区域的碱基信息，由此，可以有效地对胚胎基因组的遗传信息进行植入前检测。

此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

就本说明书而言，“计算机可读介质”可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例（非穷尽性列表）包括以下：具有一个或多个布线的电连接部（电子装置），便携式计算机盘盒（磁装置），随机存取存储器（RAM），只读存储器（ROM），可擦除可编辑只读存储器（EPROM 或闪速存储器），光纤装置，以及便携式光盘只读存储器（CDROM）。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为例如可以通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。

应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，

该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本发明各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件（例如参考 J.萨姆布鲁克等著，黄培堂等译的《分子克隆实验指南》，第三版，科学出版社）或者按照产品说明书进行。所用试剂或仪器未注明生产厂商者，均为可以通过市购获得的常规产品，例如可以采购自 Illumina 公司。

一般方法:

参考图 1，本发明实施例的主要步骤包括:

1) 获得单细胞或微量细胞样本，第三天卵裂球细胞或第五天囊胚滋养层细胞活检取样，或受精卵取样，或其他单细胞或微量细胞取样。
2) 单细胞或微量细胞样本进行全基因组扩增，扩增产物 DNA 根据新一代测序平台要求构建文库，并进行测序。
3) 单细胞或微量细胞测序数据经过过滤后与人类基因组参考序列进行比对。
4) 根据比对结果进行胚胎细胞单核苷酸多态性和缺失重复等遗传变异的初步检测。
5) 构建胚胎基因组的遗传草图，对胚胎基因型进行初始化。
6) 收集胚胎遗传相关个体，如家系成员如父母和先证者，和/或祖父母、外祖父母等或者同一对父母的其它胚胎样本，提取基因组 DNA，根据新一代测序平台要求构建文库，并进行测序。
7) 胚胎遗传相关个体 DNA 测序数据进行过滤，并与人类基因组参考序列比对。
8) 根据比对结果，分析确定先证者、父母等遗传相关个体的基因型。
9) 利用先证者、父母等遗传相关个体的基因型，推断父母双亲的单倍体型。
10) 借助隐马尔科夫模型解码方式，用父母单倍体型结果推断胚胎单体型。
11) 胚胎遗传变异的最终确定。

参考图 1，本发明实施例中信息分析部分用到的隐马尔科夫模型以及详细分析步骤如下:

标记:

1、需要检测的位点数为 N。

2、观测序列:

3、隐藏状态集为: $S=\{0,1\}$，定义了父母哪条染色体遗传给了子代。0 代表遗传给了先

证者的那条染色体。

4、观测状态集为: $V=\{0,1\}$。1 代表遗传给先证者的染色体与胚胎遗传草图基因型一致。0 代表不一致。

第一步，构建初始状态概率分布向量，单倍体重组转移矩阵以及观察序列概率矩阵

I. 初始状态概率分布记为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$

在没有参考数据的情况下，设每种隐藏状态出现的可能性相等。

II. 记单倍体重组转移矩阵记为 $A=\{a_{ij}\},(i,j\in S)$,即隐藏状态转移的概率。

$$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$$

Nr, Np 分别表示期望重组数和单核苷酸多态性位点数。

构建观察序列概率矩阵

记观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，这里的概率分布用子代一定是纯合（Must-hom）或一定是杂合(Must-het)的位点来估算。

$$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})},\ b_i(1)=1-b_i(0)$$

第二步，构建局部概率矩阵，和逆向指针

定义 局部概率 : $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$

定义 逆向指针 : $\Psi_t(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$

第三步，确定最终状态状态，并回溯最优路径

确定最终状态，$q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$

回溯最优路径，即最可能胚胎单基因型 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$

第四步，输出结果

实施例 1

一、样品收集及处理:

一对夫妇曾生育过一个重度β地中海贫血的女儿 A(先证者)，因此接受 IVF（体外受精联合胚胎移植技术）-PGD 治疗，希望能再生育一个与先证者 HLA 配型相符的健康宝宝。经过一个 IVF-PGD 治疗周期，在第三天进行卵裂球胚胎单细胞活检取样，活检后的卵裂球继续培养至第五天囊胚期，进行囊胚玻璃化冷冻保存。活检获得的卵裂球单细胞样本进行

全基因组扩增(PED7-1-OWGA)，全基因组扩增采用 Qiagen 公司的 REPLI-g MDA 扩增试剂盒，严格根据试剂盒说明书进行操作，扩增产物用于新一代测序及其他分析。同时也收集这个家系中父亲(PED4-F)、母亲(PED3-M)和先证者女儿(PED-Sister)的外周血，提取基因组 DNA。

在完成 IVF-PGD 治疗后，选择正确的胚胎植入子宫，成功妊娠后，可以在 14 周以后抽取孕妇羊水(PED7-2-OAF)，或收集胚胎脐带血，提取 DNA，进行测序，以验证胚胎细胞的分型结果。

胚胎单细胞全基因组扩增产物 DNA 用 Covaris™ 打断仪打断至 200bp 大小的片段，根据 illumia®公司 HiSeq2000™ 测序仪的上机要求进行建库，构建好的文库经 Agilent®Bioanalyzer 2100 和 Q-PCR 方法进行质控，合格的文库继续进行 illumina® HiSeq2000™ 测序仪测序，测序策略为 Pair End 90 index（即双向 90bp index 测序），测序深度 30X，全基因组测序。其中仪器的参数设置及操作方法都按照 illumina®操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

父亲、母亲和先证者女儿外周血 DNA，以及孕妇羊水或胚胎脐带血 DNA 用 Covaris™ 打断仪打断至 200bp 大小的片段，根据 illumina®公司 HiSeq2000™ 测序仪的上机要求进行建库，文库进行 All in One 芯片捕获，具体操作见 NimbleGen SeqCap EZ 芯片捕获建库说明书(http://www.nimblegen.com/products/seqcap/ez/index.html 可获取)。All in One 芯片是自主设计的目标区域捕获芯片，其目标区域包括全外显子区域、1M 的 SNP 位点区域和 MHC 基因区。芯片捕获后的文库经 Agilent®Bioanalyzer 2100 和 Q-PCR 方法进行质控，合格的文库继续进行 illumina® HiSeq2000™ 测序仪测序，测序策略为 Pair End 90 index，测序深度为目标区域的 30-50X。其中仪器的参数设置及操作方法都按照 illumina®操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

二、测序数据分析:

1、先证者和父母基因组 DNA 测序分型:

a) 过滤掉低质量值和接头污染的序列

b) 使用 Burrows-Wheeler Aligner (BWA, http://bio-bwa.sourceforge.net/)将高质量的测序序列比对在人类参考基因组（Hg19, NCBI release GRCh37)。并用 SAMtools 去除掉 PCR 重复扩增的序列。

c) 使用 Genome Analysis Toolkit (GATK, http://www.broadinstitute.org/gatk/)校准碱基质量值并进行单碱基核苷酸多态性和缺失重复的检测。

2、胚胎基因组遗传草图构建。

3、根据先证者基因型推断父母的单倍体型:

根据核心家系推断单体型，具体可参考文献 B. L. Browning and S. R. Browning. Am J Hum Genet 84:210-223.(2009)进行。

4、根据父母单体型推断胚胎细胞的单倍体型：

a) 使用 BWA 将胚胎单细胞测序数据比对到人类参考基因组（Hg19, NCBI release GRCh37)

b) 构建初始状态概率向量，以及胚胎单倍体重组转移矩阵，Nr=30，其它按上所述。

c) 统计胚胎细胞每个位点的测序信息，并构建观察序列概率矩阵（按上所述）。

d) 构建局部概率矩阵，和逆向指针（按上所述）。

e) 确定最终状态，并回溯最优路径

f) 输出

g) 以胚胎出生后的羊水基因分型结果为参照，对胚胎细胞的基因分型准确性统计如下：

种类		常染色体			性染色体		
		准确数	位点数	准确率	准确数	位点数	准确率
重构前	纯合	175,194	233,794	74.94	2,345	3,318	70.68
	杂合	81,129	81,576	99.45	1,476	1,480	99.72
重构后	纯合	172,968	173,631	99.62	2,329	2,330	99.96
	草图杂合	80,877	81,042	99.80	1,472	1,478	99.59
	更正杂合	58,194	60,194	95.88	967	990	97.68

注："草图杂合"表示重构前后均为杂合的位点；"更正杂合"表示重构前为纯合，重构后为杂合的位点。

5、β-地贫和 HLA 型别判断

a) 通过父母单体型，成功还原了胚胎细胞在 rs7480526 上的基因型，预测其为β-地贫阴性，并发现与先证者在该位置上的等位基因不一样，从而确定胚胎为β-地贫阴性。

b) 通过本方法可以得出，先证者和胚胎在 MHC 区的单体型一样，从而判断她们的 HLA 型别是匹配的。

根据上述胚胎的β-地贫基因型和 HLA 分型结果，从而选择正确的胚胎，即β-地贫阴性和 HLA 型别与先证者匹配的胚胎，进行冷冻复苏和植入子宫，从而防止患病儿的妊娠。

实施例 2

一、样品收集及处理：

一对夫妇均为β-地贫携带者，接受 IVF-PGD 治疗。经过一个 IVF-PGD 治疗周期，在第三天获得了 8 个卵裂球胚胎，均进行卵裂球胚胎单细胞活检取样，活检后的卵裂球继续培养至第五天囊胚期，进行囊胚玻璃化冷冻保存。活检获得的 8 个卵裂球单细胞样本进行全基因组扩增，全基因组扩增采用 Qiagen 公司的 REPLI-g MDA 扩增试剂盒，严格根据试剂盒说明书进行操作，扩增产物用于新一代测序及其他分析。

同时也收集这个家系中父亲、母亲的外周血，提取基因组 DNA。

在完成 IVF-PGD 治疗，选择正确的胚胎植入子宫，成功妊娠后，可以在 14 周以后抽取孕妇羊水，或收集胚胎脐带血，提取 DNA，进行测序，以验证胚胎细胞的分型结果。

8 个胚胎单细胞全基因组扩增产物 DNA，父亲、母亲外周血 DNA，以及孕妇羊水或胚胎脐带血 DNA 用 Covaris™ 打断仪打断至 200bp 大小的片段，根据 illumina®公司 HiSeq2000™ 测序仪的上机要求进行建库，文库进行 All in One 芯片捕获，具体操作见 NimbleGen SeqCap EZ 芯片捕获建库说明书 (http://www.nimblegen.com/products/seqcap/ez/index.html 可获取)。All in One 芯片是自主设计的目标区域捕获芯片，其目标区域包括全外显子区域、1M 的 SNP 位点区域和 MHC 基因区。芯片捕获后的文库经 Agilent®Bioanalyzer 2100 和 Q-PCR 方法进行质控，合格的文库继续进行 illumina® HiSeq2000™ 测序仪测序，测序策略为 Pair End 90 index（即双向 90bp index 测序），测序深度为目标区域的 30-50X。其中仪器的参数设置及操作方法都按照 illumina® 操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

二、测序数据分析:

1) 父母基因组 DNA，羊水 DNA（验证样本）测序数据分析和分型:

a) 过滤掉低质量值和接头污染的序列

b) 使用 Burrows-Wheeler Aligner (BWA, http://bio-bwa.sourceforge.net/)将高质量的测序序列比对在人类参考基因组（Hg19, NCBI release GRCh37)。并用 SAMtools 去除掉 PCR 重复扩增的序列。

c) 使用 Genome Analysis Toolkit (GATK, http://www.broadinstitute.org/gatk/)校准碱基质量值并进行单碱基核苷酸多态性和缺失重复的检测。

2) 根据 8 个胚胎单细胞测序数据分析，分别给出每个胚胎细胞基因组初始化的基因型信息，构建每个胚胎的细胞基因组遗传草图。

3）确定 8 个胚胎单细胞任一个待测胚胎，根据待测胚胎细胞基因组遗传草图和父母的基因型，对父母的单倍体型进行初始化。

4）借助隐马尔科夫模型，分别根据其余 7 个胚胎的基因型重构并确定父母的单倍体型。

5）借助隐马尔科夫模型，根据父母单倍体型，来重构待测胚胎的单体型。

6）待测胚胎遗传变异信息的确定。

以 22 号染色体为例，参照羊水基因分型结果，对待测胚胎的基因分型准确性统计如下:

类型(母 x 父)	总位点数	SNP Call 率	准确率(杂合)	准确率(纯合)
纯合 x 杂合	2183	97.16%	96.68%	99.63%
杂合 x 纯合	2156	93.18%	98.95%	100%
杂合 x 杂合	1154	82.32%	99.31%	97.59%

7) β-地贫和 HLA 型别判断

a) 通过父母单体型，成功还原了待测胚胎细胞在 rs7480526 上的单体型，预测其为 β-地贫阴性。

b) 通过本方法，结合先症者的遗传检测结果，可判定先证者和胚胎在 MHC 区的单体型是否一致，从而判断她们的 HLA 型别是否匹配。

根据上述胚胎的β-地贫基因型和 HLA 分型结果，从而选择正确的胚胎，即β-地贫阴性和 HLA 型别与先症者匹配的胚胎，进行冷冻复苏和植入子宫，从而防止患病儿的妊娠。

工业实用性

本发明的确定胚胎基因组中预定区域碱基信息的方法、用于确定胚胎基因组中预定区域碱基信息的系统以及计算机可读介质，能够有效地应用于对胚胎基因组中预定区域的核酸序列进行分析。

尽管本发明的具体实施方式已经得到详细的描述，本领域技术人员将会理解。根据已经公开的所有教导，可以对那些细节进行各种修改和替换，这些改变均在本发明的保护范围之内。本发明的全部范围由所附权利要求及其任何等同物给出。

在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示意性实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

权利要求书

1、一种确定胚胎基因组中预定区域碱基信息的方法，其特征在于，包括下列步骤：

获得胚胎细胞基因组 DNA 样本的测序结果，以及胚胎遗传相关个体基因组样本的测序结果；

基于胚胎细胞基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；

基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

2、根据权利要求 1 所述的方法，其特征在于，所述胚胎基因组 DNA 样本是从微量胚胎细胞获得的，其中，所述微量胚胎细胞来自于卵裂球、囊胚或受精卵。

3、根据权利要求 2 所述的方法，其特征在于，所述胚胎基因组 DNA 样本是从胚胎单细胞获得的。

4、根据权利要求 1 所述的方法，其特征在于，所述测序结果是利用选自 Illumina-Solexa/Hiseq、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454 和单分子测序装置的至少一种获得的。

5、根据权利要求 1 所述的方法，其特征在于，通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图。

6、根据权利要求 1 所述的方法，其特征在于，通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及

基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。

7、根据权利要求 5 或 6 所述的方法，其特征在于，所述参考序列为人类参考基因组。

8、根据权利要求 1 所述的方法，其特征在于，所述胚胎遗传相关个体包括所述胚胎的父母以及先证者。

9、根据权利要求 8 所述的方法，其特征在于，所述先证者为所述胚胎父母的其他子女。

10、根据权利要求 9 所述的方法，其特征在于，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括：

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵；

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

11、根据权利要求 10 所述的方法，其特征在于，所述隐马尔可夫模型采用下列参数：

初始状态概率分布为 $\pi = \{\pi_i\}, (i \in S, \pi_i = 0.5)$，

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 为自然数，取值范围 20-40，观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$，$b_i(1)=1-b_i(0)$，

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和；

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$，

逆向指针为 $\Psi_t(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$，

递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$，

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$ $(t=1,2,3,...,N-1)$。

12、根据权利要求 1 所述的方法，其特征在于，进一步包括：所述预定区域为已知存在遗传多态性的位点。

13、根据权利要求 12 所述的方法，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

14、一种用于确定胚胎基因组中预定区域碱基信息的系统，其特征在于，包括：

文库构建装置，所述文库构建装置适于针对胚胎基因组 DNA 样本以及胚胎遗传相关个体的基因组 DNA 样本，分别构建测序文库；

测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得胚胎以及胚胎遗传相关个体的测序结果；以及

分析装置，所述分析装置与所述测序装置相连，并且适于：

基于胚胎基因组 DNA 样本的测序结果，构建所述胚胎的遗传草图，以便确定胚胎初始基因型；

基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

15、根据权利要求 14 所述的系统，其特征在于，进一步包括 DNA 样本分离装置，所述 DNA 样本分离装置适于从胚胎进行活检取样，获得微量胚胎细胞，从所述微量胚胎细胞

提取胚胎基因组 DNA 样本，其中，所述微量胚胎细胞来自于卵裂球、囊胚或受精卵。

16、根据权利要求 14 所述的系统，其特征在于，所述测序装置为选自 Illumina-Solexa/Hiseq、ABI-SOLiD、Life Technologies-Ion Torrent/Proton、Roche-454 和单分子测序装置的至少一种。

17、根据权利要求 14 所述的系统，其特征在于，进一步包括比对装置，所述比对装置与所述测序装置相连，用于将所述测序结果与参考序列进行比对，以便:

通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图;

通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及

基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。

18、根据权利要求 14 所述的系统，其特征在于，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用惠特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

19、根据权利要求 18 所述的系统，其特征在于，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$,

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 为自然数，取值范围 20-40，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，

$b_i(0)=\frac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$, $b_i(1)=1-b_i(0)$,

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$,

逆向指针为 $\Psi_t(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$,

递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$,

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^*=\Psi_{t+1}(q^*_{t+1})$

($t=1,2,3,...,N-1$)。

20、一种计算机可读介质，其特征在于，所述计算机可读介质上存储有指令，所述指令适于被处理器执行以便:

基于胚胎细胞基因组 DNA 样本的测序结果，构建胚胎的遗传草图，以便确定胚胎初始基因型;

基于所述胚胎遗传相关个体基因组样本的测序结果，确定胚胎父母的单倍体型；以及

根据隐马尔可夫模型，以所述胚胎初始基因型作为观察序列，基于所述胚胎父母的单倍体型，确定胚胎基因组中预定区域的碱基信息。

21、根据权利要求 20 所述的计算机介质，其特征在于，通过将胚胎细胞基因组 DNA 样本的测序结果与参考序列进行比对，构建所述胚胎的遗传草图。

22、根据权利要求 20 所述的计算机介质，其特征在于，通过将所述胚胎遗传相关个体基因组样本的测序结果与参考序列进行比对，确定所述胚胎遗传相关个体的基因型；以及

基于所述胚胎遗传相关个体的基因型，确定所述胚胎父母的单倍体型。

23、根据权利要求 20 所述的计算机介质，其特征在于，根据隐马尔可夫模型，确定胚胎基因组中预定区域的碱基信息进一步包括:

构建初始状态概率分布向量、隐藏状态转移的概率矩阵以及观察序列概率矩阵;

利用患特比算法确定最终状态并回溯最优路径，以便确定胚胎基因组中预定区域的碱基信息。

24、根据权利要求 23 所述的方法，其特征在于，所述隐马尔可夫模型采用下列参数:

初始状态概率分布为 $\pi=\{\pi_i\},(i\in S,\pi_i=0.5)$，

隐藏状态转移的概率矩阵为 $A=\{a_{ij}\},(i,j\in S)$，其中，$a_{ij}=\begin{cases}1-Nr/Np & i=j\\ Nr/Np & i\neq j\end{cases}$

Nr、Np 分别表示期望重组数和单核苷酸多态性位点数，Nr 为自然数，取值范围 20-40，

观察序列概率矩阵为 $B=\{b_i(k)\},(i\in S,k\in V)$，其中，

$b_i(0)=\dfrac{\#sites(L>0,\text{Must-hom.})}{\#sites(L>0,\text{Must-het. or Must-hom.})}$，$b_i(1)=1-b_i(0)$，

#sites (L>0,Must-hom.)为子代一定为纯合的位点的数目，#sites (L>0,Must-hom. or Must-het.)为子代一定为纯合的位点的数目以及子代一定为杂合的位点的数目的总和;

局部概率为 $\delta_t(j)=\max\left[\delta_{t-1}(i)\cdot a_{ij}\right]\cdot b_j(K)\quad t\in\{1...N\}$，

逆向指针为 $\Psi_i(j)=\arg\max\delta_{t-1}(i)\cdot a_{ij}\quad t\in\{1...N\}$，

递归得最终状态为 $q_N^*=\arg\max_{q_N\in S}\delta_N(q_N)$，

回溯最优路径，确定最可能的胚胎预定区域的碱基信息为 $q_t^* = \Psi_{t+1}(q^*_{t+1})$ ($t = 1,2,3,...,N-1$)。

25、根据权利要求 20 所述的计算机可读介质，其特征在于，所述预定区域为已知存在遗传多态性的位点。

26、根据权利要求 25 所述的计算机可读介质，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

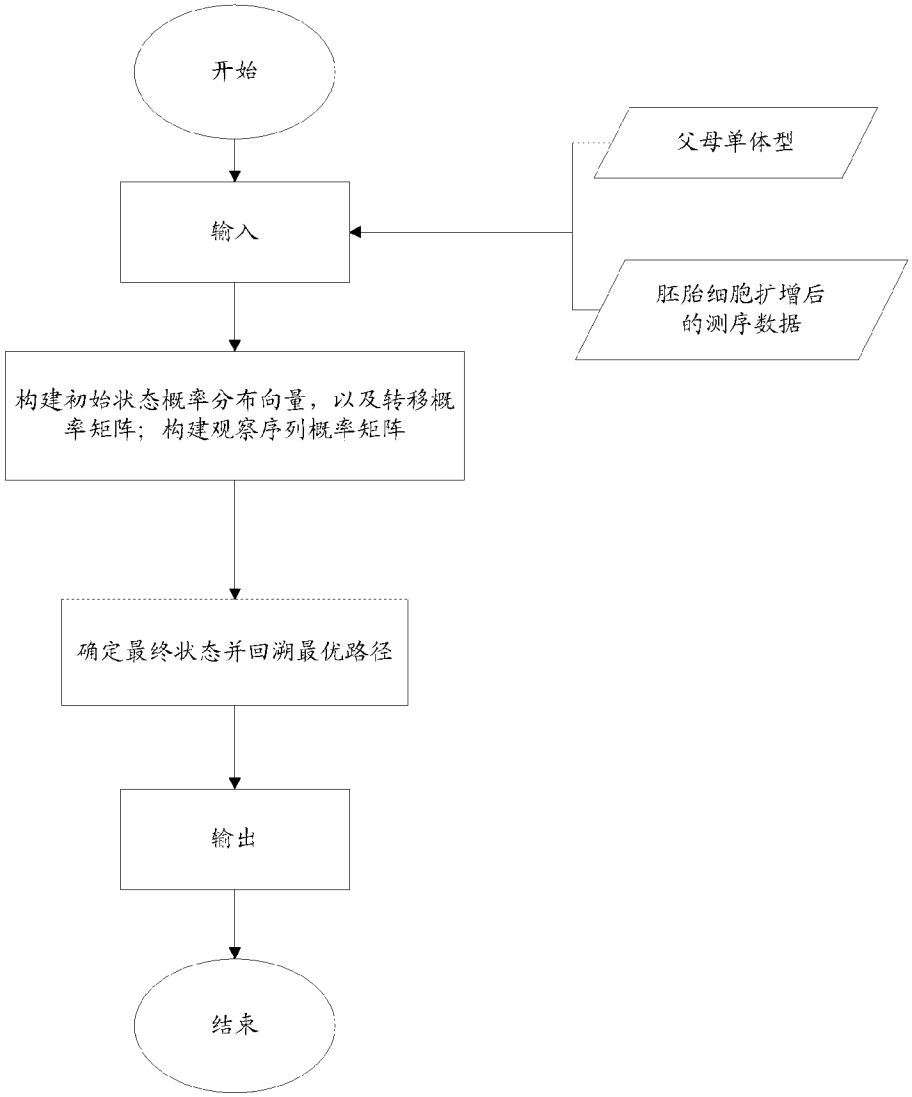

开始
输入
父母单体型
胚胎细胞扩增后的测序数据
构建初始状态概率分布向量，以及转移概率矩阵；构建观察序列概率矩阵
确定最终状态并回溯最优路径
输出
结束

图 1

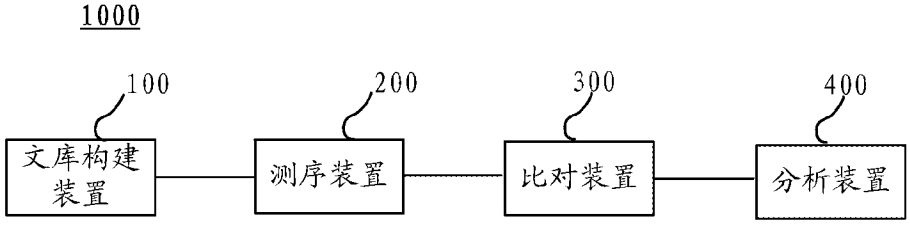

1000
100
200
300
400
文库构建装置
测序装置
比对装置
分析装置

图 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2013/073375

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i

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)

IPC: 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)

CNPAT, WPI, EPODOC, CNKI, PubMed, GOOGLE, NCBI: hidden markov model, HMM, Viterbi algorithm, library, sequencing, hidden, embryo, creat

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	YU, Weidong et al. Construction of cDNA Library from Rabbit Single Preimplantation Cloned Embryos. Chinese Journal of Biochemistry and Molecular Biology, April 2004, vol. 20, no. 2, pages 162-165, page 163, right columns 1.2.3 and 1.2.4	1-19
Y	CN 102317473 A (PACIFIC BIOSCIENCES CALIFORNIA) 11 January 2012 (11.01.2012) see claims 1 and 2, description, paragraphs [0161]-[0166]	1-19
A	CN 102127818 A (ZHANG, Kang) 20 July 2011 (20.07.2011) see the whole document	1-19

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 December 2013 (10.12.2013)	16 January 2014 (16.01.2014)

Name and mailing address of the ISA	Authorized officer
State Intellectual Property Office of the P. R. China No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088, China Facsimile No. (86-10) 62019451	ZHANG, Bin Telephone No. (86-10) 61648321

Form PCT/ISA /210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2013/073375

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.:20-26
because they relate to subject matter not required to be searched by this Authority, namely:
The subject matter of claims 20-26 relates to mere presentations of information and therefore does not warrant an international search according to the criteria set out in PCT Rule 39. 1(iv).

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2013/073375

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102317473 A	11.01.2012	WO 2010068289 A2	17.06.2010
		US 2011183320 A1	28.07.2011
		US 2010221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010068289 A3	14.10.2010
		AU 2009325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011
		US 2013029853 A1	31.01.2013
CN 102127818 A	20.07.2011	None	

Form PCT/ISA /210 (patent family annex) (July 2009)

国际检索报告

国际申请号
PCT/CN2013/073375

A. 主题的分类

C12Q 1/68 (2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

IPC : C12Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNPAT, WPI, EPODOC, CNKI, PubMed, GOOGLE, NCBI：隐马尔可夫模型，惠特比算法，文库，测序，隐马，胚胎，构建，测序，library，sequencing，hidden，HMM.

C. 相关文件

类　型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	郁卫东 等，先单个植入前克隆胚胎 cDNA 文库的构建，中国生物化学与分子生物学报，04 月 2004，第 20 卷，第 2 期，第 162-165 页 参见第 163 页右栏 1.2.3 和 1.2.4	1-19
Y	CN102317473 A（加利福尼亚太平洋生物科学股份有限公司）11.01 月 2012（11.01.2012） 参见权利要求 1-2，说明书第[0161]-[0166]段	1-19
A	CN102127818 A（张康）20.07 月 2011（20.07.2011） 参见全文	1-19

☐ 其余文件在 C 栏的续页中列出。　　☒ 见同族专利附件。

* 引用文件的具体类型：
"A" 认为不特别相关的表示了现有技术一般状态的文件
"E" 在国际申请日的当天或之后公布的在先申请或专利
"L" 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）
"O" 涉及口头公开、使用、展览或其他方式公开的文件
"P" 公布日先于国际申请日但迟于所要求的优先权日的文件
"T" 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件
"X" 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性
"Y" 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性
"&" 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
10.12 月 2013（10.12.2013）	16.1 月 2014 (16.01.2014)
ISA/CN 的名称和邮寄地址： 中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西上城路 6 号 100088 传真号：　(86-10)62019451	受权官员 张彬 电话号码：(86-10) **61648321**

PCT/ISA/210 表(第 2 页) (2009 年 7 月)

国际检索报告

国际申请号
PCT/CN2013/073375

第II栏　某些权利要求被认为是不能检索的意见(续第1页第2项)

根据条约第17条(2)(a)，对某些权利要求未做国际检索报告的理由如下：

1. ☒ 权利要求：20-26

因为它们涉及不要求本单位进行检索的主题，即：

权利要求20-26涉及单纯的信息提供（PCT细则39.1(v)）。

2. ☐ 权利要求：

因为它们涉及国际申请中不符合规定的要求的部分，以致不能进行任何有意义的国际检索，

具体地说：

3. ☐ 权利要求：

因为它们是从属权利要求，并且没有按照细则6.4(a)第2句和第3句的要求撰写。

第III栏　缺乏发明单一性的意见(续第1页第3项)

本国际检索单位在该国际申请中发现多项发明，即：

1. ☐ 由于申请人按时缴纳了被要求缴纳的全部附加检索费，本国际检索报告涉及全部可作检索的权利要求。

2. ☐ 由于无需付出有理由要求附加费的劳动即能对全部可检索的权利要求进行检索，本单位未通知缴纳任何附加费。

3. ☐ 由于申请人仅按时缴纳了部分被要求缴纳的附加检索费，本国际检索报告仅涉及已缴费的那些权利要求。

具体地说，是权利要求：

4. ☐ 申请人未按时缴纳被要求缴纳的附加检索费。因此，本国际检索报告仅涉及权利要求书中首先提及的发明；包含该发明的权利要求是：

关于异议的说明： ☐ 申请人缴纳了附加检索费，同时提交了异议书，适用时，缴纳了异议费。

☐ 申请人缴纳了附加检索费，同时提交了异议书，但未在通知书规定的时间期限内缴纳异议费。

☐ 缴纳附加检索费时未提交异议书。

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2013/073375

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN 102317473 A	11.01.2012	WO 2010068289 A2	17.06.2010
		US 2011183320 A1	28.07.2011
		US 2010221716 A1	02.09.2010
		CA 2746632 A1	17.06.2010
		WO 2010068289 A3	14.10.2010
		AU 2009325069 A1	17.06.2010
		EP 2370598 A2	05.10.2011
		US 2013029853 A1	31.01.2013
CN 102127818 A	20.07.2011	无	

Espacenet

Bibliographic data: CN104053789 (B) — 2016-02-10

Method, System And Computer Readable Medium For Determining Base Information In Predetermined Area Of Fetus Genome

Inventor(s): CHEN SHENGPEI, ; GE HUIJUAN, ; LI XUCHAO, ; YI SHANG, ; WANG JIAN, ; WANG JUN, ; YANG HUANMING, ; ZHANG XIUQIN

Applicant(s): BGI HEALTH SERVICE CO LTD

Classification: - international: *C12Q1/68*
- cooperative: **C12Q1/6809 (CN)**; **C12Q1/6858 (EP, KR, US)**; **C12Q1/6869 (EP, KR, US)**; **C12Q1/6883 (US)**; **G16B20/00 (EP, KR, US)**; **G16B30/00 (EP, KR, US)**; **G16B40/00 (KR)**; **G16H50/20 (EP, KR, US)**; **G16H50/30 (EP, KR, US)**; C12Q2600/118 (US); C12Q2600/156 (US) more

Application number: CN20128067404 20120514 Global Dossier

Priority number(s): WO2012CN75478 20120514

Also published as: CN104053789 (A) AU2012380221 (A1) AU2012380221 (B2) EP2851431 (A1) EP2851431 (A4) EP2851431 (B1) ES2656023 (T3) HK1196401 (A1) JP2015525062 (A) JP6045686 (B2) KR101770460 (B1) KR20140146193 (A) PL2851431 (T3) RU2014150655 (A) RU2597981 (C2) SG11201407515R (A) US2015094210 (A1) US2018320235 (A1) WO2013170429 (A1) less

Abstract of CN104053789 (A)

Provided are a method, system and computer readable medium for determining the base information in a predetermined area of a fetus genome, the method comprising the following steps: constructing a sequence library for the DNA samples of the fetus genome; sequencing the sequence library to obtain the sequencing result of the fetus, the sequencing result of the fetus comprised of a plurality of sequencing data; and based on the sequencing result of the fetus, determining the base information in the predetermined area according to the hidden Markov model in conjunction with the genetic information of an individual related hereditarily to the fetus.

(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 104053789 B
(45)授权公告日 2016.02.10

(21)申请号 201280067404.1

(22)申请日 2012.05.14

(85)PCT国际申请进入国家阶段日
2014.07.17

(86)PCT国际申请的申请数据
PCT/CN2012/075478 2012.05.14

(87)PCT国际申请的公布数据
WO2013/170429 ZH 2013.11.21

(73)专利权人 深圳华大基因股份有限公司
地址 518083 广东省深圳市盐田区洪安三街21号华大综合园7栋7层-14层

(72)发明人 陈盛培 葛会娟 李旭超 易赏 汪建 王俊 杨焕明 张秀清

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 李志东

(51)Int.Cl.
C12Q 1/68(2006.01)

(56)对比文件
CN 102317473 A,2012.01.11,说明书第77-161段.
CN 102127818 A,2011.07.20,说明书第8-37段.

审查员 蔡放

权利要求书3页 说明书14页 附图2页

(54)发明名称

确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

(57)摘要

提供了确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。其中,确定胎儿基因组中预定区域碱基信息的方法,包括下列步骤:针对胎儿基因组DNA样本,构建测序文库;对测序文库进行测序,以便获得胎儿的测序结果,该胎儿的测序结果由多个测序数据构成;基于胎儿的测序结果,结合胎儿遗传相关个体的遗传信息,根据隐马尔可夫模型,确定预定区域的碱基信息。

CN 104053789 B

1. 一种确定胎儿基因组中预定区域碱基信息的方法，其特征在于，包括下列步骤：

针对胎儿基因组 DNA 样本，构建测序文库；

对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；以及

基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息，所述预定区域碱基信息包括胎儿的单倍型，其中包括，

将所述胎儿的单倍型作为隐含状态，将所述胎儿的测序结果当作观察序列，借助先验数据推算出重组概率、观察序列概率分布和初始状态概率分布，按照惠特比算法，推断出最可能的所述胎儿的单体型。

2. 根据权利要求 1 所述的方法，其特征在于，所述胎儿基因组 DNA 样本是从孕妇外周血中提取的。

3. 根据权利要求 1 所述的方法，其特征在于，所述测序是利用选自 Illumina-Solexa、ABI-Solid、Roche-454 和单分子测序装置的至少一种对所述测序文库进行的。

4. 根据权利要求 1 所述的方法，其特征在于，进一步包括将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果。

5. 根据权利要求 4 所述的方法，其特征在于，所述参照序列为人类参考基因组。

6. 根据权利要求 1 所述的方法，其特征在于，所述胎儿遗传相关个体是所述胎儿的父母。

7. 根据权利要求 1 所述的方法，其特征在于，在所述惠特比算法中，采用 0.25 作为所述初始状态概率分布，采用 re/N 作为所述重组概率，其中 re = 25 ～ 30，N 为所述预定区域的长度，

采用

$$a_{jk}=\Pr\left(q_t=k|q_{t-1}=j\right)=\begin{cases}\left(1-p_r\right)^2 & x_t=x_{t-1},y_t=y_{t-1}\\ \left(1-p_r\right)\cdot p_r & x_t=x_{t-1},y_t\neq y_{t-1}\quad or\quad x_t\neq x_{t-1},y_t=y_{t-1}\\ p_r^2 & x_t\neq x_{t-1},y_t\neq y_{t-1}\end{cases}$$

作为重组转移矩阵，其中，p_r = re/N。

8. 根据权利要求 7 所述的方法，其特征在于，re = 25。

9. 根据权利要求 4 所述的方法，其特征在于，将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果进一步包括按照下列公式确定概率最高的碱基：

$$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta\left(base,m_k\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,m_{x_i}\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,f_{y_i}\right)$$

其中，

$$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

10. 根据权利要求 1 所述的方法，其特征在于，进一步包括：所述预定区域为已知存在遗传多态性的位点。

11. 根据权利要求 10 所述的方法，其特征在于，所述遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

12. 一种用于确定胎儿基因组中预定区域碱基信息的系统，其特征在于，包括：

文库构建装置，所述文库构建装置适于针对胎儿基因组 DNA 样本，构建测序文库；

测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；

分析装置，基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息，所述预定区域碱基信息包括胎儿的单倍型，其中包括，

将所述胎儿的单倍型作为隐含状态，将所述胎儿的测序结果当作观察序列，借助先验数据推算出重组概率、观察序列概率分布和初始状态概率分布，按照惠特比算法，推断出最可能的所述胎儿的单体型。

13. 根据权利要求 12 所述的系统，其特征在于，进一步包括 DNA 样本分离装置，所述 DNA 样本分离装置适于从孕妇外周血中提取胎儿基因组 DNA 样本。

14. 根据权利要求 12 所述的系统，其特征在于所述测序装置为选自 Illumina-Solexa、ABI-Solid、Roche-454 和单分子测序装置的至少一种。

15. 根据权利要求 12 所述的系统，其特征在于，进一步包括比对装置，所述比对装置与所述测序装置相连，用于将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果。

16. 根据权利要求 12 所述的系统，其特征在于，所述惠特比算法采用 0.25 作为所述初始状态概率分布，采用 re/N 作为所述重组概率，其中 re = 25 ～ 30，N 为所述预定区域的长度，

采用

$$a_{jk}=\Pr\left(q_i=k\middle|q_{i-1}=j\right)=\begin{cases}\left(1-p_r\right)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ \left(1-p_r\right)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\ \ or\ \ x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

作为重组转移矩阵，其中，p_r = re/N。

17. 根据权利要求 16 所述的系统，其特征在于，re = 25。

18. 根据权利要求 15 所述的系统，其特征在于，将所述胎儿的测序结果与参照序列进行比对，以便确定来自于所述预定区域的测序结果进一步包括按照下列公式确定概率最高的碱基：

$$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta\left(base,m_k\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,m_{x_i}\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,f_{y_i}\right)$$

其中，

$$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质

[0001] 优先权信息

[0002] 无

技术领域

[0003] 本发明涉及确定胎儿基因组中预定区域碱基信息的方法、系统和计算机可读介质。

背景技术

[0004] 遗传性疾病是由于遗传物质发生改变而造成的疾病，具有先天性、家族性、终身性和遗传性的特点。遗传性疾病可分为3个大类：单基因遗传病、多基因遗传病及染色体异常。其中单基因病多由于单个致病基因的显性或隐性遗传所致基因功能异常；而多基因遗传病则是由多个基因变化影响所致的疾病，会在一定程度上受到外界环境因素的影响；染色体异常包括数目异常和结构异常，最为多见的是由于第21号染色体三体所致的唐氏综合症，患儿表现为先天愚型和肢体形状异常等其他先天性特征。由于目前对遗传性疾病尚无有效的治疗方式，只能针对性地进行支持治疗或者药物缓解，费用昂贵，给社会和家庭带来沉重经济和精神负担。因此，在孩子出生前就对孩子的患病状态进行检测，做好预防工作，以达到优生优育的目的，是十分必要的。

[0005] 然而，目前的相关检测手段仍有待改进。

发明内容

[0006] 本发明旨在至少解决现有技术中存在的技术问题之一。

[0007] 在本发明的一个方面，本发明提出了一种确定胎儿基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：针对胎儿基因组DNA样本，构建测序文库；对所述测序文库进行测序，以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。对于孕期血浆，假若我们将胎儿的单倍型（父母单倍型的重组型）作为隐含状态(hidden states)，可以将血浆的测序数据当做观察序列(observations)，借助先验数据推算出状态转移概率(transition probabilities)、观察序列概率分布(observation symbol probabilities)和初始状态概率分布(initial state distribution)，我们则可以通过诸如惠特比算法(Viterbi algorithm)根据隐马尔可夫模型推断出最可能的胎儿单倍型组合，从而获得更多胎儿的信息。因而，根据本发明的实施例，借助隐马尔可夫模型，例如可以通过利用惠特比算法(Viterbi algorithm)，参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0008] 在本发明的又一方面，本发明提出了一种用于确定胎儿基因组中预定区域碱基信息的系统。根据本发明的实施例，该系统包括：文库构建装置，所述文库构建装置适于针对胎儿基因组 DNA 样本，构建测序文库；测序装置，所述测序装置与所述文库构建装置相连，并且适于对所述测序文库进行测序，以便以便获得胎儿的测序结果，所述胎儿的测序结果由多个测序数据构成；分析装置，基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。利用该系统，能够有效地实施前面所述的确定胎儿基因组中预定区域碱基信息的方法，可以借助隐马尔可夫模型，例如可以通过利用惠特比算法 (Viterbi algorithm)，参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测，从而可以有效地对胎儿基因组的遗传信息进行产前确定。

[0009] 在本发明的另一方面，本发明还提出了一种计算机可读介质。根据本发明的实施例，该计算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定预定区域的碱基信息。利用本发明的计算机可读介质，能够有效地被处理器执行其存储的指令，以便借助隐马尔可夫模型，例如可以通过利用惠特比算法 (Viterbi algorithm)，基于胎儿的测序结果，参考胎儿遗传相关个体的遗传信息，可以确定胎儿基因组中特定区域的核酸序列，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0010] 本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0011] 本发明的上述和 / 或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

[0012] 图 1 为根据本发明一个实施例的利用隐马尔可夫模型进行分析的流程示意图；以及

[0013] 图 2 为根据本发明的一个实施例的用于确定胎儿基因组中预定区域核酸序列的系统的结构示意图。

具体实施方式

[0014] 下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，仅用于解释本发明，而不能理解为对本发明的限制。

[0015] 需要说明的是，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征。进一步地，在本发明的描述中，除非另有说明，“多个”的含义是两个或两个以上。

[0016] 确定胎儿基因组中预定区域碱基信息的方法

[0017] 在本发明的第一方面，本发明提出了一种确定胎儿基因组中预定区域碱基信息的方法。根据本发明的实施例，该方法包括下列步骤：

[0018] 首先，针对胎儿基因组 DNA 样本，构建测序文库。根据本发明的实施例，胎儿基因组 DNA 样本的来源并不受特别限制。根据本发明的一些实施例，可以采用任何含有胎儿核酸的孕妇样本。例如，根据本发明的实施例，可以采用的孕妇样本为孕妇乳汁、尿液和外周血。其中，优选孕妇外周血。采用孕妇外周血作为胎儿基因组 DNA 样本的来源，可以有效地实现无创取样方式获得胎儿基因组 DNA，从而可以在不影响胎儿正常发育的前提下，对胎儿的基因组进行有效监测。关于针对核酸样本，构建测序文库的方法和流程，本领域技术人员可以根据不同的测序技术进行适当选择，关于流程的细节，可以参见测序仪器的厂商例如 Illumina 公司所提供的规程，例如参见 Illumina 公司 Multiplexing Sample Preparation Guide (Part#1005361 ;Feb2010) 或 Paired-End SamplePrep Guide(Part#1005063 ; Feb2010)，通过参照将其并入本文。根据本发明的实施例，从生物样本提取核酸样本的方法和设备，也不受特别限制，可以采用商品化的核酸提取试剂盒进行。

[0019] 在构建测序文库后，将测序文库应用于测序仪器，对测序文库进行测序，并获得相应的测序结果，该测序结果是由多个测序数据构成的。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，能够提高后续对测序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台 (MetzkerML. Sequencing technologies-the next generation. Nat Rev Genet. 2010Jan ;11(1):31-46) 包括但不限于 Illumina-Solexa(GA™, HiSeq2000™等)、ABI-Solid 和 Roche-454(焦磷酸测序) 测序平台；单分子测序平台 (技术) 包括但不限于 Helicos 公司的真实单分子测序技术 (True Single Molecule DNA sequencing)，Pacific Biosciences 公司单分子实时测序 (single molecule real-time(SMRT™))，以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等 (Rusk, Nicole(2009-04-01). Cheap Third-Generation Sequencing. Nature Methods6(4):244-245)。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。

[0020] 任选地，在得到测序结果之后，可以将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据。在本文中所使用的术语“预定区域”应作广义理解，是指任何包含可能发生预定事件位点的核酸分子的区域。对于 SNP 分析而言，可以是指包含 SNP 位点的区域。对于分析染色体非整倍性，则预定区域指的是所要分析的染色体的全长或者部分，即选择所有来自该染色体的测序数据。从测序结果中选择来自相应区域的测序数据的方法可以不受特别限制。根据本发明的实施例，可以通过将所得到的所有测序数据与已知的核酸参照序列进行比对，从而得到来自于预定区域的测序数据。另外，根据本发明的实施例，预定区域也可以是基因组上不连续的多个分散点。根据本发明的实施例，可以使用的参照序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参照序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI36.3,HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

[0021] 接下来，基于与预定区域对应的测序数据，确定预定区域中的部分核酸序列；以及基于所确定的预定区域中的部分核酸序列，按照惠特比算法，确定预定区域的其他核酸序列，以便获得预定区域的核酸序列。根据本发明的实施例，可以通过基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，可以利用惠特比算法 (Viterbi algorithm)，借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息。由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0022] 下面参考图 1，对利用惠特比算法借助隐马尔可夫模型进行分析的原理进行详细描述：

[0023] 在本文中所使用的术语“胎儿遗传相关个体”指的是在遗传意义上，与胎儿之间具有亲缘关系的个体，例如根据本发明的实施例，可以采用的“胎儿遗传相关个体”为胎儿的亲代例如父母。由此，子代的基因组形成过程，相当于亲代基因组的一次随机重组（即连锁互换单倍体型重组，以及配子的随机组合）。对于孕期血浆，假若我们将胎儿的单倍型（父母单倍型的重组型）作为隐含状态 (hidden states)，可以将血浆的测序数据当做观察序列 (observations)，借助先验数据推算出状态转移概率 (transition probabilities)、观察序列概率分布 (observation symbol probabilities) 和初始状态概率分布 (initial state distribution)，我们则可以通过惠特比算法 (Viterbi algorithm) 推断出最可能的胎儿单倍型组合，从而获得更多胎儿的信息。

[0024] 详细分析步骤如下：

[0025] 记号：

[0026] I. 需要检测的位点数为 N。

[0027] II. 父母的单倍型分别记为 $FH = \{fh_0, fh_1\}$ 和 $MH = \{mh_0, mh_1\}$，

[0028] 其中，

[0029] $mh_k = \{m_{1,k}, \ldots, m_{i,k}, \ldots, m_{N,k}\}$，$fh_k = \{f_{1,k}, \ldots, f_{i,k}, \ldots, f_{N,k}\}$，

[0030] $\forall fh_{i,k}, mh_{i,k} \in \{A, C, G, T\}$，

[0031] $k \in \{0,1\}$，$i = 1, 2, 3, \ldots, N$。

[0032] III. 将未知的胎儿单倍型记为 $H = \{h_0, h_1\}$，特别地，h_0 和 h_1 分别遗传自母亲和父亲。

[0033] $h_0 = \{m_{1,x_1}, \ldots, m_{i,x_i}, \ldots, m_{N,x_N}\}$，$h_1 = \{f_{1,y_1}, \ldots, f_{i,y_i}, \ldots, f_{N,y_N}\}$，

[0034] 其中，$x_i \in \{0,1\}$，$y_i \in \{0,1\}$，

[0035] 下标 x_i 和 y_i 组成的序列对，$q_i = \{x_i, y_i\}$ 组成了我们需要解码的隐藏状态，

[0036] 而所有可能出现的隐藏状态组成了集合 Q。

[0037] IV. 测序数据记为，$S = \{s_1, \ldots, s_i, \ldots, s_N\}$，

[0038] 其中，$s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$，代表此位点的测序信息，包含了 ACGT 四种碱基的数量。

[0039] V. 平均胎儿浓度和平均测序错误率分别记为 ε 和 e。

[0040] 第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵：

[0041] I. 初始状态概率分布记为 $\pi = \{\pi_j\}$ $(j \in Q)$。

[0042] 根据本发明的实施例，在没有参考数据的情况下，可以设$\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$，即每种隐藏状态在第一个位点出现的可能性相等。

[0043] II. 根据本发明的实施例，记单倍体重组概率为$p_r = re/N$，其中 re 代表人类配子基因组重组平均次数，为先验数据在 25 到 30 之间。

[0044] III. 根据本发明的实施例，记单倍体重组转移矩阵记为$A = \{a_{jk}\}(j,k \in Q)$，其中$a_{jk}$为隐藏状态转移的概率，即

[0045] $$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r)\cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases},$$

[0046] 胎儿单倍型$h_0 = \{m_{1,x_1}, ..., m_{i,x_i}, ..., m_{N,x_N}\}$，$h_1 = \{f_{1,y_1}, ..., f_{i,y_i}, ..., f_{N,y_N}\}$的下标$x_i$和$y_i$组成的序列对，$q_i = \{x_i, y_i\}$组成了我们需要解码的隐藏状态。举个例子，$x_i = 0$代表，“母源性染色体上，对应基因座上等位基因型为$m_{i,0}$”。

[0047] 第二步，构建观察序列概率矩阵：

[0048] 根据本发明的实施例，记观察序列概率矩阵为$B = \{b_{i,j}(s_i)\}(i = 1,2,3,...,N, j \in Q)$，其中$b_{i,j}(s_i)$代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j，$j = \{x_i, y_i\}$）时，观测到这种测序信息的可能性”，即

[0049] $$b_{i,j}(s_i) = \Pr(s_i | q_i = j, \{m_0, m_1\}) = \frac{(n_{i,A}+n_{i,C}+n_{i,G}+n_{i,T})!}{n_{i,A}!n_{i,C}!n_{i,G}!n_{i,T}!} \cdot (P_{i,A})^{n_{i,A}} \cdot (P_{i,C})^{n_{i,C}} \cdot (P_{i,G})^{n_{i,G}} \cdot (P_{i,T})^{n_{i,T}},$$

[0050] 其中$P_{i,base}$代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j，$j = \{x_i, y_i\}$）时，该碱基出现的可能性”，即

[0051] $$P_{i,base} = \Pr(base | q_i = j, \{m_0, m_1\}) = \sum_{k \in \{0,1\}} \frac{1}{2}(1-\varepsilon)\Delta(base, m_k) + \frac{1}{2}\varepsilon \cdot \Delta(base, m_{x_i}) + \frac{1}{2}\varepsilon \cdot \Delta(base, f_{y_i}),$$

[0052] 其中指示函数

[0053] $$\Delta(x,y) = \begin{cases} 1-e & x = y \\ e/3 & x \neq y \end{cases}.$$

[0054] 此步是进行 HMM 参数，每个位点的观察序列概率分布$b_{i,j}(s_i)$计算，即计算每个位点上不同胎儿单倍型（隐藏状态）下，血浆出现当前测序数据（观察序列）的可能性。

[0055] 第三步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）：

[0056] 定义局部概率$\delta_i(q_i) = \left(\max_{q_{i-1} \in Q} \delta_i(q_i) \cdot a_{q_{i-1}q_i}\right) \cdot b_{i,q_i}(s_i)$

[0057] 定义逆向指针 $\Psi_i(q_i)=\arg\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}$

[0058] 这里所使用的术语局部概率 $\delta_i(q_i)$ 和逆向指针 $\Psi_i(q_i)$ 都是沿用 Viterbi 算法的经典定义。关于该参数的定义的详细描述，可以参见 LawrenceR. Rabiner, PROCEEDINGS OF THE IEEE, Vol. 77, No. 2, 1989 年 2 月，通过参照将其全文并入本文。

[0059] 第四步，确定最终状态，并回溯最优路径：

[0060] 确定最终状态，$q_N^*=\arg\max_{q_N\in Q}\delta_N(q_N)$

[0061] 按照逆向指针回溯最优路径，即最可能胎儿单基因型 $q_i^*=\Psi_i(q_i)\ (i=1,2,3,...,N-1)$。

[0062] 第五步，输出结果。

[0063] 由此，能够有效地对胎儿基因组的序列进行分析。相比其他已有的产前检测技术方法，本方法有以下技术优势，主要体现在准确性和可获得的遗传信息量上：

[0064] 1) 根据本发明实施例，检测的位点不仅针对父源性位点，对母源性位点，即母亲杂合位点，同样可以很好地检测出来胎儿是否遗传了母亲的致病位点，检测准确率可高达 95％以上，且可以检测多种变异类型，扩大了疾病检测的范围。

[0065] 2) 根据本发明实施例，不但可以通过一次测序获得多个位点、多种疾病的信息，对于一些在孕妇血浆中覆盖程度较低，单纯提高测序深度依然无法准确判定的基因序列，可以通过本方法推断得到，结果准确可靠。

[0066] 3) 根据本发明实施例，可进行遗传疾病作图，对于一些连锁相关疾病，可通过其他位点的信息直接推断出来，一次可获得的信息量大，对临床检测更加具有指导意义。

[0067] 另外，根据本发明实施例，本发明的确定胎儿基因组中预定区域碱基信息的方法，不仅限于 SNP 或者 STR 等某一种遗传多态性位点，对所有的遗传多态性位点均可适用，且可以多种位点同时使用，以便互相验证。除了可进行产前无创检测胎儿基因组信息，达到疾病检测的目的，还可以进行无创产前亲子鉴定，在孩子出生前判定孩子父亲身份，为一些涉及抚养责任和义务、财产纠纷、性侵案等协助侦破。

[0068] 用于确定胎儿基因组中预定区域碱基信息的系统

[0069] 在本发明的又一方面，本发明提出了一种用于确定胎儿基因组中预定区域核酸序列的系统。根据本发明的实施例，参考图 2，该系统 1000 可以包括：文库构建装置 100、测序装置 200 以及分析装置 400。

[0070] 根据本发明的实施例，文库构建装置 100 适于针对胎儿基因组 DNA 样本，构建测序文库。根据本发明的实施例，测序装置 200 与文库构建装置 100 相连，并且适于对所构建的测序文库进行测序，以便获得测序结果，所得到的测序结果由多个测序数据构成。根据本发明的实施例，还可以进一步包括 DNA 样本分离装置，该 DNA 样本分离装置适于从孕妇外周血中提取胎儿基因组 DNA 样本。由此，该系统可以适用于进行无创产前检测。

[0071] 根据本发明的实施例，任选地，还可以包括比对装置 300。根据本发明的实施例，比对装置 300 与测序装置 200 相连，并且适于将所得到的测序结果与参照序列进行比对，以便确定与预定区域对应的测序数据。根据本发明的实施例，可以用于进行测序的方法和设备并不受特别限制，包括但不限于双脱氧链终止法；优选高通量的测序方法，由此，能够利用这些测序装置的高通量、深度测序的特点，进一步提高测序效率。从而，提高后续对测

序数据进行分析，尤其是统计检验分析时的精确性和准确度。所述高通量的测序方法包括但不限于第二代测序技术或者是单分子测序技术。所述第二代测序平台 (MetzkerML. Sequencing technologies-the next generation. Nat Rev Genet. 2010Jan ;11(1):31-46) 包括但不限于 Illumina-Solexa(GA™, HiSeq2000™等)、ABI-Solid 和 Roche-454(焦磷酸测序)测序平台；单分子测序平台（技术）包括但不限于 Helicos 公司的真实单分子测序技术 (True Single Molecule DNA sequencing)，Pacific Biosciences 公司单分子实时测序 (single molecule real-time(SMRT™))，以及 Oxford Nanopore Technologies 公司的纳米孔测序技术等 (Rusk, Nicole(2009-04-01). Cheap Third-Generation Sequencing. Nature Methods6(4):244-245)。随着测序技术的不断进化，本领域技术人员能够理解的是还可以采用其他的测序方法和装置进行全基因组测序。根据本发明的具体示例，可以利用选自 Illumina-Solexa、ABI-SOLiD、Roche-454 和单分子测序装置的至少一种对所述全基因组测序文库进行测序。根据本发明的实施例，可以使用的参照序列的类型并不受特别限制，可以为任何含有感兴趣区域的已知序列。根据本发明的实施例，可以采用已知的人类参考基因组作为参照序列。例如，根据本发明的实施例，采用的人类参考基因组为 NCBI36.3，HG18。另外，根据本发明的实施例，进行比对的方法并不受特别限制。根据本发明的具体实施例，可以采用 SOAP 进行比对。

[0072] 根据本发明的实施例，分析装置 400 适于基于所述胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。

[0073] 根据本发明的实施例，惠特比算法采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re = 25 ～ 30，优选 25，N 为所述预定区域的长度，

[0074] 采用

[0075] $$a_{jk}=\Pr(q_i=k|q_{i-1}=j)=\begin{cases}(1-p_r)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ (1-p_r)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\ \ or\ \ x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

[0076] 作为重组转移矩阵，p_r = re/N。

[0077] 根据本发明的实施例，将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据进一步包括按照下列公式确定概率最高的碱基：

[0078] $$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta(base,m_k)+\frac{1}{2}\varepsilon\cdot\Delta(base,m_{x_i})+\frac{1}{2}\varepsilon\cdot\Delta(base,f_{y_i})$$

[0079] 其中，

[0080] $$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

[0081] 关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胎儿基因组中预定区域核酸序列的系统。不再赘述。

[0082] 由此，利用该系统，能够有效地实施前面所述的确定胎儿基因组中预定区域核酸序列的方法，可以通过例如惠特比算法 (Viterbi algorithm)，借助隐马尔可夫模型，确定

胎儿基因组中特定区域的碱基信息，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0083] 此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

[0084] 在本文中所述的术语“相连”应作广义理解，既可以是直接相连，也可以是间接相连，只要能够实现上述功能上的衔接即可。

[0085] 需要说明的是，本领域技术人员能够理解，在前面所描述的确定胎儿基因组中预定区域核酸序列的方法的特征和优点也适合于确定胎儿基因组中预定区域核酸序列的系统，为描述方便，不再详述。

[0086] 计算机可读介质

[0087] 在本发明的又一方面，本发明提出了一种计算机可读介质。根据本发明的实施例，计算机可读介质上存储有指令，所述指令适于被处理器执行以便基于胎儿的测序结果，结合胎儿遗传相关个体的遗传信息，根据隐马尔可夫模型，确定所述预定区域的碱基信息。由此，利用该计算机可读介质，能够有效地实施前面所述的方法，从而可以通过例如惠特比算法 (Viterbi algorithm)，借助隐马尔可夫模型，确定胎儿基因组中特定区域的碱基信息，由此，可以有效地对胎儿基因组的遗传信息进行产前检测。

[0088] 根据本发明的实施例，指令适于按照惠特比算法，根据隐马尔可夫模型，确定所述预定区域的碱基信息。根据本发明的实施例，在所述惠特比算法中，采用 0.25 作为初始状态概率分布，采用 re/N 作为重组概率，其中 re = 25 ～ 30，优选 25，N 为所述预定区域的长度，

[0089] 采用

[0090] $$a_{jk}=\Pr(q_i=k|q_{i-1}=j)=\begin{cases}(1-p_r)^2 & x_i=x_{i-1},y_i=y_{i-1}\\ (1-p_r)\cdot p_r & x_i=x_{i-1},y_i\neq y_{i-1}\ \ or\ \ x_i\neq x_{i-1},y_i=y_{i-1}\\ p_r^2 & x_i\neq x_{i-1},y_i\neq y_{i-1}\end{cases}$$

[0091] 作为重组转移矩阵，其中，p_r = re/N。

[0092] 根据本发明的实施例，所述指令将所述测序结果与参照序列进行比对，以便确定与所述预定区域对应的测序数据进一步包括按照下列公式确定概率最高的碱基：

[0093] $$P_{i,base}=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta(base,m_k)+\frac{1}{2}\varepsilon\cdot\Delta(base,m_{x_i})+\frac{1}{2}\varepsilon\cdot\Delta(base,f_{y_i})$$

[0094] 其中，

[0095] $$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}$$

。

[0096] 关于数据分析部分，前面已经进行了详细描述，也当然地适用于确定胎儿基因组中预定区域核酸序列的系统。不再赘述。

[0097] 此外，根据本发明的实施例，预定区域为已知存在遗传多态性的位点，而遗传多态性为选自单核苷酸多态性和 STR 的至少一种。

[0098] 就本说明书而言，“计算机可读介质”可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例（非穷尽性列表）包括以下：具有一个或多个布线的电连接部（电子装置），便携式计算机盘盒（磁装置），随机存取存储器（RAM），只读存储器（ROM），可擦除可编辑只读存储器（EPROM或闪速存储器），光纤装置，以及便携式光盘只读存储器（CDROM）。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为例如可以通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。

[0099] 应当理解，本发明的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。例如，如果用硬件来实现，和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

[0100] 本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

[0101] 此外，在本发明各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用时，也可以存储在一个计算机可读取存储介质中。

[0102] 下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件（例如参考J.萨姆布鲁克等著，黄培堂等译的《分子克隆实验指南》，第三版，科学出版社）或者按照产品说明书进行。所用试剂或仪器未注明生产厂商者，均为可以通过市购获得的常规产品，例如可以采购自Illumina公司。

[0103] 一般方法

[0104] 本发明实施例的主要步骤包括：

[0105] 1）无创采取含有胎儿遗传物质的孕妇样品，提取其中含有的遗传物质。

[0106] 2）胎儿家庭成员如父母和外祖父母等基因组DNA提取和纯化。

[0107] 3）各遗传物质根据不同测序平台测序要求进行文库构建。

[0108] 4）测序获得的数据进行过滤，过滤条件根据质量值、接头污染等来设定。

[0109] 5）获得的高质量序列根据需要进行组装处理，组装结果与人类基因组参考序列进行比对。获得唯一比对的序列，带入模型进行分析。

[0110] 分析模型：

[0111] 记号：

[0112] I.需要检测的位点数为N。

[0113] II. 父母的单倍型分别记为 $FH = \{fh_0, fh_1\}$ 和 $MH = \{mh_0, mh_1\}$，

[0114] 其中，

[0115] $mh_k = \{m_{1,k}, \ldots, m_{i,k}, \ldots, m_{N,k}\}$，$fh_k = \{f_{1,k}, \ldots, f_{i,k}, \ldots, f_{N,k}\}$，

[0116] $\forall fh_{i,k}, mh_{i,k} \in \{A, C, G, T\}$，

[0117] $k \in \{0, 1\}$，$i = 1, 2, 3, \ldots, N$。

[0118] III. 将未知的胎儿单倍型记为 $H = \{h_0, h_1\}$，特别地，h_0 和 h_1 分别遗传自母亲和父亲。

[0119] $h_0 = \{m_{1,x_1}, \ldots, m_{i,x_i}, \ldots, m_{N,x_N}\}$，$h_1 = \{f_{1,y_1}, \ldots, f_{i,y_i}, \ldots, f_{N,y_N}\}$，

[0120] 其中，$x_i \in \{0, 1\}$，$y_i \in \{0, 1\}$，

[0121] 下标 x_i 和 y_i 组成的序列对，$q_i = \{x_i, y_i\}$ 组成了我们需要解码的隐藏状态，

[0122] 而所有可能出现的隐藏状态组成了集合 Q。

[0123] IV. 测序数据记为，$S = \{s_1, \ldots, s_i, \ldots, s_N\}$

[0124] 其中，$s_i = \{n_{i,A}, n_{i,C}, n_{i,G}, n_{i,G}\}$，代表此位点的测序信息，包含了 ACGT 四种碱基的数量。

[0125] V. 平均胎儿浓度和平均测序错误率分别记为 ε 和 e。

[0126] 第一步，构建初始状态概率分布向量，以及单倍体重组转移矩阵：

[0127] I. 初始状态概率分布记为 $\pi = \{\pi_j\}(j \in Q)$，

[0128] 根据本发明的实施例，在没有参考数据的情况下，可以设 $\pi_j = \Pr(q_1 = j) \triangleq \frac{1}{4}$，即每种隐藏状态在第一个位点出现的可能性相等。

[0129] II. 根据本发明的实施例，记单倍体重组概率为 $p_r = re/N$，其中 re 代表人类配子基因组重组平均次数，为先验数据在 25 到 30 之间。

[0130] III. 根据本发明的实施例，记单倍体重组转移矩阵记为 $A = \{a_{jk}\}(j, k \in Q)$，其中 a_{jk} 为隐藏状态转移的概率，即

[0131] $$a_{jk} = \Pr(q_i = k | q_{i-1} = j) = \begin{cases} (1-p_r)^2 & x_i = x_{i-1}, y_i = y_{i-1} \\ (1-p_r) \cdot p_r & x_i = x_{i-1}, y_i \neq y_{i-1} \quad or \quad x_i \neq x_{i-1}, y_i = y_{i-1} \\ p_r^2 & x_i \neq x_{i-1}, y_i \neq y_{i-1} \end{cases}$$

[0132] 第二步，构建观察序列概率矩阵：

[0133] 根据本发明的实施例，记观察序列概率矩阵为 $B = \{b_{i,j}(s_i)\}(i = 1, 2, 3, \ldots, N, j \in Q)$，其中 $b_{i,j}(s_i)$ 代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j）时，观测到这种测序信息的可能性”，即

[0134] $$b_{i,j}(s_i) = \Pr(s_i | q_i = j, \{m_0, m_1\}) = \frac{(n_{i,A} + n_{i,C} + n_{i,G} + n_{i,T})!}{n_{i,A}! n_{i,C}! n_{i,G}! n_{i,T}!} \cdot (P_{i,A})^{n_{i,A}} \cdot (P_{i,C})^{n_{i,C}} \cdot (P_{i,G})^{n_{i,G}} \cdot (P_{i,T})^{n_{i,T}}$$

[0135] 其中 $P_{i,base}$ 代表“在位点 i，考虑母亲单倍型和胎儿单倍型（状态 j）时，该碱基出

现的可能性”，即

[0136] $$P_{i,base}=\Pr\left(base\middle|q_i=j,\{m_0,m_1\}\right)=\sum_{k\in\{0,1\}}\frac{1}{2}(1-\varepsilon)\Delta\left(base,m_k\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,m_{x_i}\right)+\frac{1}{2}\varepsilon\cdot\Delta\left(base,f_{y_i}\right),$$

[0137] 其中指示函数

[0138] $$\Delta(x,y)=\begin{cases}1-e & x=y\\ e/3 & x\neq y\end{cases}。$$

[0139] 第三步，构建局部概率矩阵，和逆向指针（下面以一维局部概率矩阵构建为例）：

[0140] 定义局部概率 $\delta_i(q_i)=\left(\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}\right)\cdot b_{i,q_i}(s_i)$

[0141] 定义逆向指针 $\Psi_i(q_i)=\arg\max_{q_{i-1}\in Q}\delta_i(q_i)\cdot a_{q_{i-1}q_i}$

[0142] 第四步，确定最终状态状态，并回溯最优路径：

[0143] 确定最终状态，$q_N^*=\arg\max_{q_N\in Q}\delta_N(q_N)$，

[0144] 根据逆向指针回溯最优路径，即最可能胎儿单基因型 $q_i^*=\Psi_i(q_i)\ (i=1,2,3,\ldots,N-1)$。

[0145] 第五步，输出结果。

[0146] 实施例 1

[0147] 样品收集及处理：

[0148] (1) 所取样品包括一个家庭中父亲和母亲孕期的外周血，胎儿出生后取脐带血，以EDTA 抗凝管收集，Oragene® DNA 唾液采集 /DNA 纯化试剂盒 OG-250 采集祖父母和外祖父母唾液。

[0149] (2) 祖父母和外祖父母唾液 DNA 提取后用 Infinium® HD Human610-Quad BeadChip 基因芯片进行分型。

[0150] (3) 取母亲孕期外周血，1600g，4℃离心 10 分钟，将血细胞和血浆分开，血浆再以 16000g，4℃离心 10 分钟，进一步去除残留的白细胞。孕妇血浆用 TIANamp Micro DNA Kit(TIANGEN) 提取 DNA，得到母亲和胎儿基因组 DNA 混合物，并且从全血分离血浆后剩余的白细胞中提取母亲基因组 DNA。将所得到的血浆 DNA 根据 Illumia® 公司 HiSeq2000™ 测序仪的上机要求进行建库，构建好的文库经 Agilent® Bioanalyzer2100 检测片段分布范围符合要求，再经过 Q-PCR 方法对两个文库进行定量，合格后 Illumina® HiSeq2000™ 测序仪测序，测序循环数为 PE101index（即双向 101bp index 测序），其中仪器的参数设置及操作方法都按照 Illumina® 操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

[0151] (4) 父亲外周血、母亲外周血白细胞和胎儿脐带血则直接用 TIANamp Micro DNA Kit(TIANGEN) 提取试剂盒提取基因组 DNA。

[0152] 除血浆DNA样品外，将所获得的所有DNA样品，需用Covaris™打断仪打断至500bp大小的片段。将获得的DNA片段以及血浆DNA样品根据Illumia®公司HiSeq2000™测序仪的上机要求进行建库，具体步骤如下：

[0153] 末端修复反应体系：

[0154]

10×T4 多核苷酸激酶缓冲液	10 μl
dNTPs(10mM)	4 μl
T4 DNA 聚合酶	5 μl
Klenow 片段	1 μl
T4 多核苷酸激酶	5 μl
DNA 片段	30μl
ddH_2O 补齐至 100 μl	

[0155] 20℃反应30分钟后，使用PCR Purification Kit(QIAGEN)回收末端修复产物。将所得到的产物最后溶于34μl的EB缓冲液中。

[0156] 末端添加碱基A反应体系：

[0157]

10×Klenow 缓冲液	5μl
dATP(1mM)	10μl
Klenow (3'-5' exo$^-$)	3μl
DNA	32μl

[0158] 37℃温育30分钟后，经MinElute® PCR Purification Kit(QIAGEN)纯化并溶于12μl的EB中。

[0159] 接头连接反应体系：

[0160]

2x Rapid DNA 连接缓冲液	25μl
PEI Adapter oligomix(20uM)	10μl
T4 DNA 连接酶	5μl
添加碱基A的产物	10μl

[0161] 20℃反应15分钟后，使用PCR Purification Kit(QIAGEN)回收连接产物。将所得到的产物最后溶于32μl的EB缓冲液中。

[0162]

PCR 反应体系：	
接头连接反应产物	10 μl
Phusion DNA Polymerase Mix	25 μl
PCR 引物 (10 pmol/μl)	1 μl
Index N(10 pmol/μl)	1 μl
超纯水	13 μl

[0163] 反应程序如下：

[0164]

98℃	30 s	
98℃	10 s	10 个循环
65℃	30 s	
72℃	30 s	
72℃	5 min	
4℃	Hold	

[0165] 使用 PCR Purification Kit(QIAGEN) 回收 PCR 产物。样品最后溶于 50 μ l 的 EB 缓冲液中。

[0166] 构建好的文库经 Agilent® Bioanalyzer2100 检测片段分布范围符合要求，再经过 Q-PCR 方法对两个文库进行定量，合格后，用 Illumina® HiSeq2000™测序仪测序，测序循环数为 PE101index（即双向 101bp index 测序），其中仪器的参数设置及操作方法都按照 Illumina® 操作手册（可由 http://www.illumina.com/support/documentation.ilmn 获取）。

[0167] (5) 父母基因组测序分型：

[0168] a. 使用 SOAP2 将测序数据比对到人类参考基因组（版本为 NCBI36.3，HG18）。

[0169] b. 使用 SOAPsnp（南方汉族 (CHS) 家系数据使用的是千人计划数据）对数据进行一致序列 (consensus sequence，CNS) 构建。

[0170] c. 提取出标记位点的基因型。

[0171] (6) 父母单倍体型推断：

[0172] a. 构建含祖辈与父母基因型的群体基因型矩阵，即提取父母、祖辈和南方汉族家系在标记位点的基因型。

[0173] b. 使用 BEAGLE 对父母的单倍型进行推断。

[0174] (7) 胎儿单倍体型推断：

[0175] a. 用 SOAP2 将血浆测序数据比对到人类参考基因组（版本为 NCBI36.3，HG18）。

[0176] b. 构建初始状态概率向量，以及单倍体重组转移矩阵。

[0177] 构建初始状态概率向量：采取无参考数据模式，即各个初始状态概率相等，均为 0.25。

[0178] 单倍体重组转移矩阵：保守地，我们取 re = 25（其余按一般方法所述）。

[0179] c. 统计每个位点的测序信息，并构建观察序列概率矩阵（其余按一般方法所述）。

Pg. 1141

[0180] d. 构建局部概率矩阵，和逆向指针（其余按一般方法所述）。

[0181] e. 确定最终状态状态，并回溯最优路径。

[0182] f. 输出。

[0183] 根据胎儿出生后的脐带血基因分型结果，我们的分类准确性统计如下：

[0184]

			母亲						合计		
			纯合			杂合					
			位点数	准确数	准确率	位点数	准确数	准确率	位点数	准确数	准确率
常染色体	父亲	纯合	199,552	199,552	100.00%	66,238	63,988	96.57%	265,790	263,520	99.15%
		杂合	65,409	64,735	98.97%	41,849	39,944	95.45%	107,258	104,679	97.60%
	合计		264,961	264,287	99.75%	108,087	103,912	96.14%	373,048	368,199	98.70%
X染色体			4,881	4,881	100.00%	1,718	1,478	86.03%	6,599	6,359	96.36%

[0185] 工业实用性

[0186] 本发明的确定胎儿基因组中预定区域碱基信息的方法、用于确定胎儿基因组中预定区域碱基信息的系统以及计算机可读介质，能够有效地应用于对胎儿基因组中预定区域的核酸序列进行分析。

[0187] 尽管本发明的具体实施方式已经得到详细的描述，本领域技术人员将会理解。根据已经公开的所有教导，可以对那些细节进行各种修改和替换，这些改变均在本发明的保护范围之内。本发明的全部范围由所附权利要求及其任何等同物给出。

[0188] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示意性实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不一定指的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任何的一个或多个实施例或示例中以合适的方式结合。

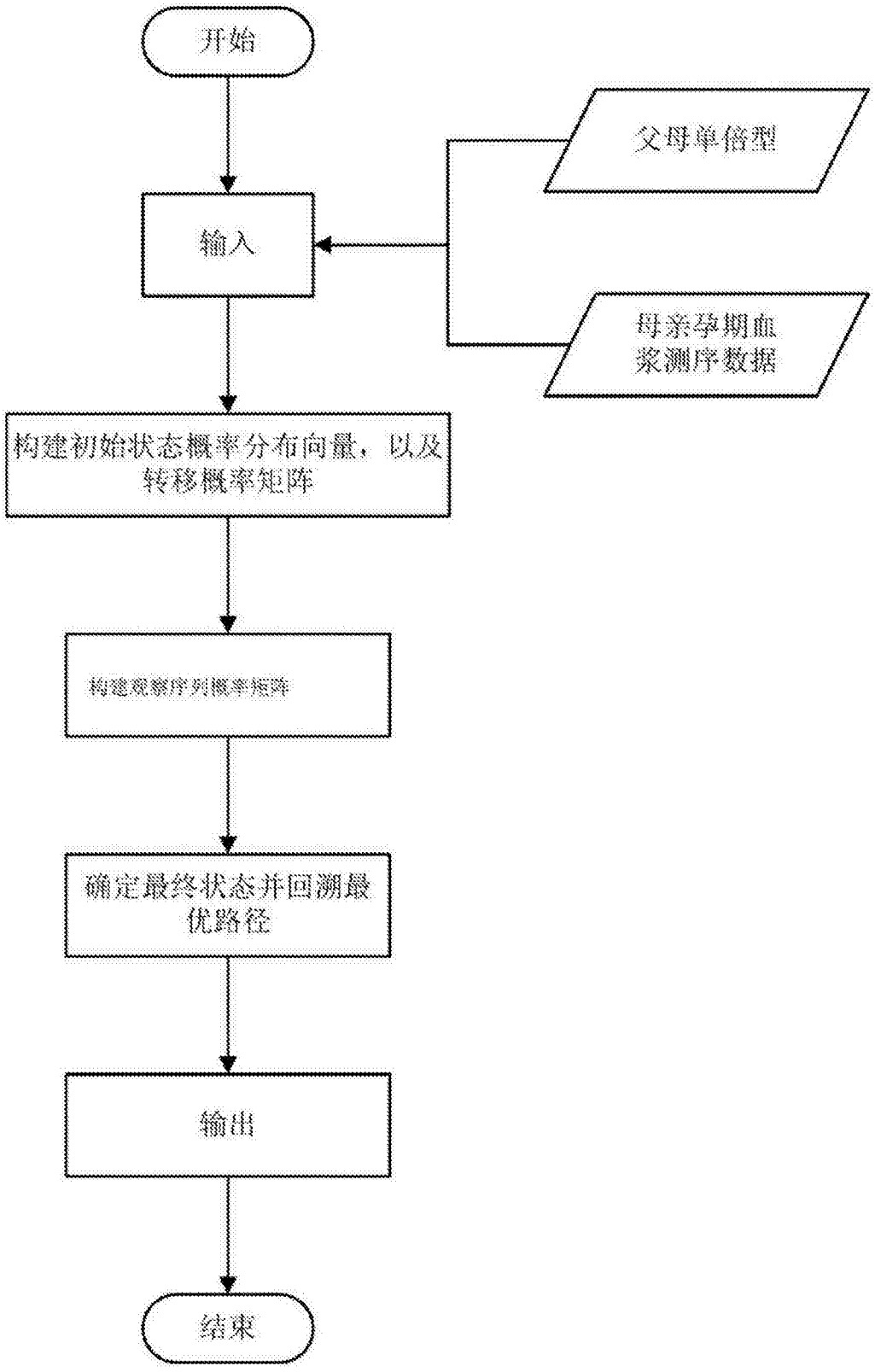

开始
输入
父母单倍型
母亲孕期血浆测序数据
构建初始状态概率分布向量，以及转移概率矩阵
构建观察序列概率矩阵
确定最终状态并回溯最优路径
输出
结束

图 1

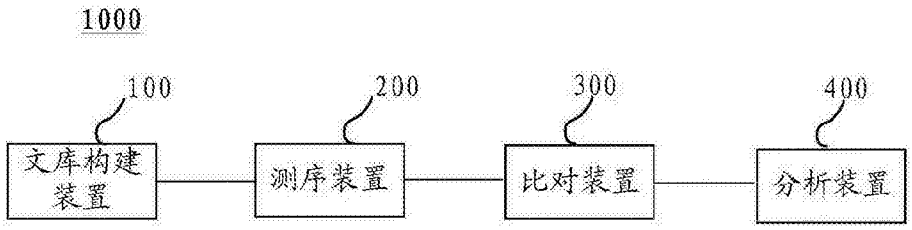


图 2

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 January 2009 (29.01.2009)

PCT

(10) International Publication Number
WO 2009/013492 A1

(51) International Patent Classification:
G06F 19/00 (2006.01) *C12Q 1/68* (2006.01)

(21) International Application Number:
PCT/GB2008/002524

(22) International Filing Date: 23 July 2008 (23.07.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/951,438 23 July 2007 (23.07.2007) US

(71) Applicant *(for all designated States except US)*: **THE CHINESE UNIVERSITY OF HONG KONG** [CN/CN]; Technology Licensing Office, Room 226, Pi Ch'iu Building, Shatin, New Territories (HK).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **LO, Yuk-Ming Dennis** [GB/CN]; 4th Floor, 7 King Tak Street, Homantin, Kowloon (HK). **CHIU, Rossa Wai Kwun** [AU/CN]; Flat 1A, Block 1, Constellation Cove, 1 Hung Lam Drive, Tai Po, New Territories (HK). **CHAN, Kwan Chee** [CN/CN]; Flat A, 13/F, Block 34, Broadway Street, Mei Foo Sun Chuen, Kowloon (HK). **ZEE, Benny Chung Ying** [CA/CN]; Flat 18E, Tower 2, La Costa, Ma on Shan, New Territories (HK). **CHONG, Ka Chun** [CN/CN]; Flat 06, 29/F, Shin King House, Fu Shin Estate, Tai Po, New Territories (HK).

(74) Agents: **BUFTON, Karen** et al.; Mewburn Ellis LLP, 33 Gutter Lane, London EC2V 8AS (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT,

[Continued on next page]

(54) Title: DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE

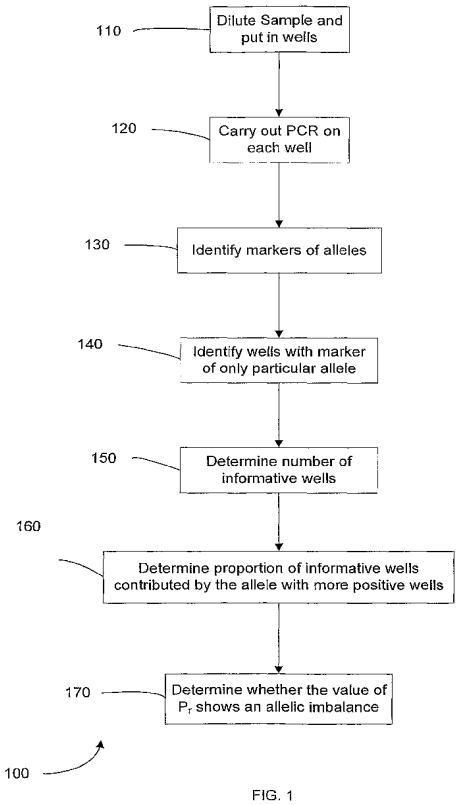


FIG. 1

(57) Abstract: Methods, systems, and apparatus are provided for determining whether a nucleic acid sequence imbalance exists within a biological sample. One or more cutoff values for determining an imbalance of, for example, the ratio of the two sequences (or sets of sequences) are chosen. The cutoff value may be determined based at least in part on the percentage of fetal DNA in a sample, such as maternal plasma, containing a background of maternal nucleic acid sequences. The cutoff value may also be determined based on an average concentration of a sequence per reaction. In one aspect, the cutoff value is determined from a proportion of informative wells that are estimated to contain a particular nucleic acid sequence, where the proportion is determined based on the above-mentioned percentage and/or average concentration. The cutoff value may be determined using many different types of methods, such as sequential probability ratio testing (SPRT).

WO 2009/013492 A1

RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE

CLAIM OF PRIORITY

[0001] The present application claims priority from and is a non-provisional application of U.S. Provisional Application No. 60/951438, entitled "DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE" filed July 23, 2007 (Attorney Docket No. 016285-005200US), the entire contents of which are herein incorporated by reference for all purposes.

CROSS-REFERENCES TO RELATED APPLICATIONS

[0002] The present application is also related to concurrently filed non-provisional application entitled "DIAGNOSING FETAL CHROMOSOMAL ANEUPLOIDY USING GENOMIC SEQUENCING," (Attorney Docket No. 016285-005220US), the entire contents of which are herein incorporated by reference for all purposes.

FIELD OF THE INVENTION

[0003] This invention generally relates to the diagnostic testing of genotypes and diseases by determining an imbalance between two different nucleic acid sequences, and more particularly to the identification of Down syndrome, other chromosomal aneuploidies, mutations and genotypes in a fetus via testing a sample of maternal blood. The invention also relates to the detection of cancer, the monitoring of transplantation, and the monitoring of infectious diseases.

BACKGROUND

[0004] Genetic diseases, cancers, and other conditions often result from or produce an imbalance in two corresponding chromosomes or alleles or other nucleic acid sequences. That is an amount of one sequence relative to another sequence is larger or smaller than normal. Usually, the normal ratio is an even 50/50 ratio. Down Syndrome (trisomy 21) is such a disease having an imbalance of an extra chromosome 21.

[0005] Conventional prenatal diagnostic methods of trisomy 21 involve the sampling of fetal materials by invasive procedures such as amniocentesis or chorionic villus sampling, which pose a finite risk of fetal loss. Non-invasive procedures, such as screening by ultrasonography and biochemical markers, have been used to risk-stratify pregnant women prior to definitive invasive diagnostic procedures. However, these screening methods typically measure epiphenomena that are associated with trisomy 21 instead of the core

chromosomal abnormality, and thus have suboptimal diagnostic accuracy and other disadvantages, such as being highly influenced by gestational age.

[0006] The discovery of circulating cell-free fetal DNA in maternal plasma in 1997 offered new possibilities for noninvasive prenatal diagnosis (Lo, YMD and Chiu, RWK 2007 *Nat Rev Genet* 8, 71-77). While this method has been readily applied to the prenatal diagnosis of sex-linked (Costa, JM et al. 2002 *N Engl J Med* 346, 1502) and certain single gene disorders (Lo, YMD et al. 1998 *N Engl J Med* 339, 1734-1738), its application to the prenatal detection of fetal chromosomal aneuploidies has represented a considerable challenge (Lo, YMD and Chiu, RWK 2007, supra). First, fetal nucleic acids co-exist in maternal plasma with a high background of nucleic acids of maternal origin that can often interfere with the analysis (Lo, YMD et al. 1998 *Am J Hum Genet* 62, 768-775). Second, fetal nucleic acids circulate in maternal plasma predominantly in a cell-free form, making it difficult to derive dosage information of genes or chromosomes within the fetal genome.

[0007] Significant developments overcoming these challenges have recently been made (Benachi, A & Costa, JM 2007 *Lancet* 369, 440-442). One approach detects fetal-specific nucleic acids in the maternal plasma, thus overcoming the problem of maternal background interference (Lo, YMD and Chiu, RWK 2007, supra). Dosage of chromosome 21 was inferred from the ratios of polymorphic alleles in the placenta-derived DNA/RNA molecules. However, this method is less accurate when samples contain lower amount of the targeted gene and can only be applied to fetuses who are heterozygous for the targeted polymorphisms, which is only a subset of the population if one polymorphism is used.

[0008] Dhallan et al (Dhallan, R, *et al.* 2007, supra Dhallan, R, *et al.* 2007 *Lancet* 369, 474-481) described an alternative strategy of enriching the proportion of circulating fetal DNA by adding formaldehyde to maternal plasma. The proportion of chromosome 21 sequences contributed by the fetus in maternal plasma was determined by assessing the ratio of paternally-inherited fetal-specific alleles to non-fetal-specific alleles for single nucleotide polymorphisms (SNPs) on chromosome 21. SNP ratios were similarly computed for a reference chromosome. An imbalance of fetal chromosome 21 was then inferred by detecting a statistically significant difference between the SNP ratios for chromosome 21 and those of the reference chromosome, where significant is defined using a fixed p-value of ≤ 0.05. To ensure high population coverage, more than 500 SNPs were targeted per chromosome. However, there have been controversies regarding the effectiveness of formaldehyde to

enrich to a high proportion (Chung, GTY, et al. 2005 *Clin Chem* 51, 655-658), and thus the reproducibility of the method needs to be further evaluated. Also, as each fetus and mother would be informative for a different number of SNPs for each chromosome, the power of the statistical test for SNP ratio comparison would be variable from case to case (Lo, YMD & Chiu, RWK. 2007 *Lancet* 369, 1997). Furthermore, since these approaches depend on the detection of genetic polymorphisms, they are limited to fetuses heterozygous for these polymorphisms.

[0009] Using polymerase chain reaction (PCR) and DNA quantification of a chromosome 21 locus and a reference locus in amniocyte cultures obtained from trisomy 21 and euploid fetuses, Zimmermann et al (2002 *Clin Chem* 48, 362-363) were able to distinguish the two groups of fetuses based on the 1.5-fold increase in chromosome 21 DNA sequences in the former. Since a 2-fold difference in DNA template concentration constitutes a difference of only one threshold cycle (Ct), the discrimination of a 1.5-fold difference has been the limit of conventional real-time PCR. To achieve finer degrees of quantitative discrimination, alternative strategies are needed. Accordingly, some embodiments of the present invention use digital PCR (Vogelstein, B et al. 1999 *Proc Natl Acad Sci U S A* 96, 9236-9241) for this purpose.

[0010] Digital PCR has been developed for the detection of allelic ratio skewing in nucleic acid samples (Chang, HW *et al.* 2002 *J Natl Cancer Inst* 94, 1697-1703). Clinically, it has been shown to be useful for the detection of loss of heterozygosity (LOH) in tumor DNA samples (Zhou, W. *et al.* 2002 *Lancet* 359, 219-225). For the analysis of digital PCR results, sequential probability ratio testing (SPRT) has been adopted by previous studies to classify the experimental results as being suggestive of the presence of LOH in a sample or not (El Karoui at al. 2006 *Stat Med* 25, 3124-3133). In methods used in the previous studies, the cutoff value to determine LOH used a fixed reference ratio of the two alleles in the DNA of 2/3. As the amount, proportion, and concentration of fetal nucleic acids in maternal plasma are variable, these methods are not suitable for detecting trisomy 21 using fetal nucleic acids in a background of maternal nucleic acids in maternal plasma.

[0011] It is desirable to have a noninvasive test for fetal trisomy 21 (and other imbalances) detection based on circulating fetal nucleic acid analysis, especially one that is independent of the use of genetic polymorphisms and/or of fetal-specific markers. It is also desirable to have accurate determination of cutoff values and counting of sequences, which can reduce the

number of wells of data and/or the amount of maternal plasma nucleic acid molecules necessary for accuracy, thus providing increased efficiency and cost-effectiveness. It is also desirable that noninvasive tests have high sensitivity and specificity to minimize false diagnoses.

[0012] Another application for fetal DNA detection in maternal plasma is for the prenatal diagnosis of single gene disorders such as beta-thalassemia. However, as fetal DNA only constitutes a minor fraction of DNA in maternal plasma, it is thought that this approach can only detect a mutation that a fetus has inherited from its father, but which is absent from the mother. Examples of this include the 4 bp deletion in codon 41/42 of the beta-globin gene causing beta-thalassemia (Chiu RWK et al. 2002 *Lancet*, 360, 998-1000) and the Q890X mutation of the cystic fibrosis transmembrance conductance regulator gene causing cystic fibrosis (Gonzalez-Gonzalez et al 2002 *Prenat Diagn,* 22, 946-8). However, as both beta-thalassemia and cystic fibrosis are autosomal recessive conditions, in which the fetus would need to inherit a mutation from each parent before the disease would manifest itself, the detection of merely the paternally-inherited mutation would only increase the risk of having the fetus having the disease from 25% to 50%. Diagnostically this is not ideal. Thus, the main diagnostic application of the existing approach would be for the scenario when no paternally-inherited fetal mutation can be detected in maternal plasma, when the fetus can then be excluded from having the homozygous disease state. However, diagnostically, this approach has the disadvantage that the conclusion is made based on the negative detection of the paternal mutation. Thus, an approach which would allow the complete fetal genotype (be it homozygous normal, homozygous mutant, or heterozygous) to be determined from maternal plasma, without the above limitation, would be very desirable.

BRIEF SUMMARY

[0013] Embodiments of this invention provides methods, systems, and apparatus for determining whether a nucleic acid sequence imbalance (e.g., allelic imbalance, mutational imbalance, or chromosome imbalance) exists within a biological sample. One or more cutoff values for determining an imbalance of, for example, a ratio of amounts of the two sequences (or sets of sequences) are chosen.

[0014] In one embodiment, the cutoff value is determined based at least in part on the percentage of fetal (clinically relevant nucleic acid) sequences in a biological sample, such as maternal plasma or serum or urine, which contains a background of maternal nucleic acid

sequences. In another embodiment, the cutoff value is determined based on an average concentration of a sequence in a plurality of reactions. In one aspect, the cutoff value is determined from a proportion of informative wells that are estimated to contain a particular nucleic acid sequence, where the proportion is determined based on the above-mentioned percentage and/or average concentration.

[0015] The cutoff value may be determined using many different types of methods, such as SPRT, false discovery, confidence interval, receiver operating characteristic (ROC). This strategy further minimized the amount of testing required before confident classification could be made. This is of particular relevance to plasma nucleic acid analysis where the template amount is often limiting.

[0016] According to one exemplary embodiment, a method is provided for determining whether a nucleic acid sequence imbalance exists within a biological sample, the method comprising: receiving data from a plurality of reactions, wherein the data includes: (1) a first set of quantitative data indicating a first amount of a clinically relevant nucleic acid sequence; and (2) a second set of quantitative data indicating a second amount of a background nucleic acid sequence different from the clinically relevant nucleic acid sequence; determining a parameter from the two data sets; deriving a first cutoff value from an average concentration of a reference nucleic acid sequence in each of the plurality of reactions, wherein the reference nucleic acid sequence is either the clinically relevant nucleic acid sequence or the background nucleic acid sequence; comparing the parameter to the first cutoff value; and based on the comparison, determining a classification of whether a nucleic acid sequence imbalance exists.

[0017] According to another exemplary embodiment, a method is provided for determining whether a nucleic acid sequence imbalance exists within a biological sample, the method comprising: receiving data from a plurality of reactions, wherein the data includes: (1) a first set of quantitative data indicating a first amount of a clinically relevant nucleic acid sequence; and (2) a second set of quantitative data indicating a second amount of a background nucleic acid sequence different from the clinically relevant nucleic acid sequence, wherein the clinically relevant nucleic acid sequence and the background nucleic acid sequence come from a first type of cells and from one or more second types of cells; determining a parameter from the two data sets; deriving a first cutoff value from a first percentage resulting from a measurement of an amount of a nucleic acid sequence from the first type of cells in the

biological sample; comparing the parameter to the cutoff value; and based on the comparison, determining a classification of whether a nucleic acid sequence imbalance exists.

[0018] Other embodiments of the invention are directed to systems and computer readable media associated with methods described herein.

[0019] A better understanding of the nature and advantages of the present invention may be gained with reference to the following detailed description and the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0020] FIG. 1 is a flowchart illustrating a digital PCR experiment.

[0021] FIG. 2A illustrates a digital RNA-SNP and RCD method according to an embodiment of the present invention.

[0022] FIG. 2B shows a table of examples of frequently detectable chromosomal aberrations in cancers.

[0023] FIG. 3 illustrates a graph having SPRT curves used to determine Down syndrome according to an embodiment of the present invention.

[0024] FIG. 4 shows a method of determining a disease state using a percentage of fetal cells according to an embodiment of the present invention.

[0025] FIG. 5 shows a method of determining a disease state using an average concentration according to an embodiment of the present invention.

[0026] FIG. 6 shows a table that tabulates the expected digital RNA-SNP allelic ratio and P_r of trisomy 21 samples for a range of template concentrations expressed as the average reference template concentration per well (m_r) according to an embodiment of the present invention.

[0027] FIG. 7 shows a table that tabulates the expected P_r for the fractional fetal DNA concentrations of 10%, 25%, 50% and 100% in trisomy 21 samples at a range of template concentrations expressed as the average reference template concentration per well (m_r) according to an embodiment of the present invention.

[0028] FIG. 8 shows a plot illustrating the degree of differences in the SPRT curves for m_r values of 0.1, 0.5 and 1.0 for digital RNA-SNP analysis according to an embodiment of the present invention.

[0029] FIG. 9A shows a table of a comparison of the effectiveness of the new and old SPRT algorithms for classifying euploid and trisomy 21 cases in 96-well digital RNA-SNP analyses according to an embodiment of the present invention.

[0030] FIG. 9B shows a table of a comparison of the effectiveness of the new and old SPRT algorithms for classifying euploid and trisomy 21 cases in 384-well digital RNA-SNP analyses according to an embodiment of the present invention.

[0031] FIG. 10 is a table showing the percentages of fetuses correctly and incorrectly classified as euploid or aneuploid and those not classifiable for the given informative counts according to an embodiment of the present invention.

[0032] FIG. 11 is a table 1100 showing computer simulations for digital RCD analysis for a pure (100%) fetal DNA sample according to an embodiment of the present invention.

[0033] FIG. 12 is a table 1200 showing results of computer simulation of accuracies of digital RCD analysis at m_r=0.5 for the classification of samples from euploid or trisomy 21 fetuses with different fractional concentrations of fetal DNA according to an embodiment of the present invention.

[0034] FIG. 13A shows a table 1300 of digital RNA-SNP analysis in placental tissues of euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0035] FIG. 13B shows a table 1350 of digital RNA-SNP analysis of maternal plasma from euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0036] FIG. 14A-14C show plots illustrating a cutoff curve resulting from an RCD analysis according to an embodiment of the present invention.

[0037] FIG. 15A shows a table of digital RNA-SNP analysis in placental tissues of euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0038] FIG. 15B shows a table of digital RNA-SNP data of the 12 reaction panels from one maternal plasma sample according to an embodiment of the present invention.

[0039] FIG. 15C shows a table of digital RNA-SNP analysis of maternal plasma from euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0040] FIG. 16A shows a table for a digital RNA-SNP analysis of euploid and trisomy 18 placentas according to an embodiment of the present invention.

[0041] FIG. 16B shows an SPRT interpretation of digital RNA-SNP data for euploid and trisomy 18 placentas according to an embodiment of the present invention.

[0042] FIG. 17 shows a table of a digital RCD analysis of 50% placental/maternal blood cell DNA mixtures of euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0043] FIG. 18 shows a SPRT curve illustrating the decision boundaries for correct classification according to an embodiment of the present invention.

[0044] FIG. 19 shows a table of digital RCD analysis of amniotic fluid samples from euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

[0045] FIG. 20 shows a table of digital RCD analysis of placental DNA samples from euploid and trisomy 18 pregnancies (E= euploid; T18= trisomy 18) according to an embodiment of the present invention.

[0046] FIG. 21 shows a table of a multiplex digital RCD analysis of 50% placental/maternal blood cell DNA mixtures of euploid and trisomy 21 pregnancies (E= euploid; T21= trisomy 21; U= unclassified) according to an embodiment of the present invention.

[0047] FIGS. 22A and 22B show a table of a multiplex digital RCD analysis of 50% euploid or trisomy 21 placental genomic DNA/50% maternal buffy coat DNA mix according to an embodiment of the present invention. Unclass denotes unclassifiable and T21 denotes trisomy 21.

[0048] FIG. 23 shows a scenario where both the male and female partners carry the same mutation.

[0049] FIG. 24A shows a table of a digital RMD analysis of female/male and male/male DNA mixtures according to an embodiment of the present invention.

[0050] FIG. 24B shows a table of a digital RMD analysis of mixtures with 25% female and 75% male DNA according to an embodiment of the present invention.

[0051] FIG. 25 shows a table of a digital RMD analysis of 15%-50% DNA mixtures mimicking maternal plasma samples for HbE mutation according to an embodiment of the present invention.

[0052] FIG. 26A shows a table of a digital RMD analysis of 5%-50% DNA mixtures mimicking maternal plasma samples for CD41/42 mutation according to an embodiment of the present invention.

[0053] FIG. 26B shows a table of a digital RMD analysis of 20% DNA mixtures mimicking maternal plasma samples for CD41/42 mutation according to an embodiment of the present invention.

[0054] FIG. 27 shows a block diagram of an exemplary computer apparatus usable with system and methods according to embodiments of the present invention.

DEFINITIONS

[0055] The term "*biological sample*" as used herein refers to any sample that is taken from a subject (*e.g.*, a human, such as a pregnant woman) and contains one or more nucleic acid molecule(s)of interest.

[0056] The term "*nucleic acid*" or "*polynucleotide*" refers to a deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) and a polymer thereof in either single- or double-stranded form. Unless specifically limited, the term encompasses nucleic acids containing known analogs of natural nucleotides that have similar binding properties as the reference nucleic acid and are metabolized in a manner similar to naturally occurring nucleotides. Unless otherwise indicated, a particular nucleic acid sequence also implicitly encompasses conservatively modified variants thereof (*e.g.*, degenerate codon substitutions), alleles, orthologs, SNPs, and complementary sequences as well as the sequence explicitly indicated. Specifically, degenerate codon substitutions may be achieved by generating sequences in which the third position of one or more selected (or all) codons is substituted with mixed-base and/or deoxyinosine residues (Batzer *et al.*, *Nucleic Acid Res.* **19**:5081 (1991); Ohtsuka *et al.*, *J. Biol. Chem.* **260**:2605-2608 (1985); and Rossolini *et al.*, *Mol. Cell. Probes* **8**:91-98 (1994)). The term nucleic acid is used interchangeably with gene, cDNA, mRNA, small noncoding RNA, micro RNA (miRNA), Piwi-interacting RNA, and short hairpin RNA (shRNA) encoded by a gene or locus.

[0057] The term "*gene*" means the segment of DNA involved in producing a polypeptide chain. It may include regions preceding and following the coding region (leader and trailer) as well as intervening sequences (introns) between individual coding segments (exons).

[0058] The term "*reaction*" as used herein refers to any process involving a chemical, enzymatic, or physical action that is indicative of the presence or absence of a particular polynucleotide sequence of interest. An example of a "*reaction*" is an amplification reaction such as a polymerase chain reaction (PCR). Another example of a "*reaction*" is a sequencing reaction, either by synthesis or by ligation. An "*informative reaction*" is one that indicates the presence of one or more particular polynucleotide sequence of interest, and in one case where only one sequence of interest is present. The term "*well*" as used herein refers to a reaction at a predetermined location within a confined structure, e.g., a well-shaped vial, cell, or chamber in a PCR array.

[0059] The term "*clinically relevant nucleic acid sequence*" as used herein can refer to a polynucleotide sequence corresponding to a segment of a larger genomic sequence whose potential imbalance is being tested or to the larger genomic sequence itself. One example is the sequence of chromosome 21. Other examples include chromosome 18, 13, X and Y. Yet other examples include mutated genetic sequences or genetic polymorphisms or copy number variations that a fetus may inherit from one or both of its parents. Yet other examples include sequences which are mutated, deleted, or amplified in a malignant tumor, e.g. sequences in which loss of heterozygosity or gene duplication occur. In some embodiments, multiple clinically relevant nucleic acid sequences, or equivalently multiple makers of the clinically relevant nucleic acid sequence, can be used to provide data for detecting the imbalance. For instance, data from five non-consecutive sequences on chromosome 21 can be used in an additive fashion for the determination of a possible chromosomal 21 imbalance, effectively reducing the need of sample volume to 1/5.

[0060] The term "*background nucleic acid sequence*" as used herein refers to a nucleic acid sequence whose normal ratio to the clinically relevant nucleic acid sequence is known, for instance a 1-to-1 ratio. As one example, the background nucleic acid sequence and the clinically relevant nucleic acid sequence are two alleles from the same chromosome that are distinct due to heterozygosity. In another example, the background nucleic acid sequence is one allele that is heterozygous to another allele that is the clinically relevant nucleic acid sequence. Moreover, some of each of the background nucleic acid sequence and the clinically relevant nucleic acid sequence may come from different individuals.

[0061] The term "*reference nucleic acid sequence*" as used herein refers to a nucleic acid sequence whose average concentration per reaction is known or equivalently has been measured.

[0062] The term "*overrepresented nucleic acid sequence*" as used herein refers to the nucleic acid sequence among two sequences of interest (e.g., a clinically relevant sequence and a background sequence) that is in more abundance than the other sequence in a biological sample.

[0063] The term "*based on*" as used herein means "*based at least in part on*" and refers to one value (or result) being used in the determination of another value, such as occurs in the relationship of an input of a method and the output of that method. The term "*derive*" as used herein also refers to the relationship of an input of a method and the output of that method, such as occurs when the derivation is the calculation of a formula.

[0064] The term "*quantitative data*" as used herein means data that are obtained from one or more reactions and that provide one or more numerical values. For example, the number of wells that show a fluorescent marker for a particular sequence would be quantitative data.

[0065] The term "*parameter*" as used herein means a numerical value that characterizes a quantitative data set and/or a numerical relationship between quantitative data sets. For example, a ratio (or function of a ratio) between a first amount of a first nucleic acid sequence and a second amount of a second nucleic acid sequence is a parameter.

[0066] The term "*cutoff value*" as used herein means a numerical value whose value is used to arbitrate between two or more states (e.g. diseased and non-diseased) of classification for a biological sample. For example, if a parameter is greater than the cutoff value, a first classification of the quantitative data is made (e.g. diseased state); or if the parameter is less than the cutoff value, a different classification of the quantitative data is made (e.g. non-diseased state).

[0067] The term "*imbalance*" as used herein means any significant deviation as defined by at least one cutoff value in a quantity of the clinically relevant nucleic acid sequence from a reference quantity. For example, the reference quantity could be a ratio of 3/5, and thus an imbalance would occur if the measured ratio is 1:1.

DETAILED DESCRIPTION OF THE INVENTION

[0068] This invention provides methods, systems, and apparatus for determining whether an increase or decrease compared to a reference (e.g. non-diseased) quantity of a clinically-relevant nucleic acid sequence in relation to other non-clinically-relevant sequences (e.g., a chromosomal or allelic imbalance) exists within a biological sample. One or more cutoff values are chosen for determining whether a change compared to the reference quantity exists (i.e. an imbalance), for example, with regards to the ratio of amounts of two sequences (or sets of sequences). The change detected in the reference quantity may be any deviation (upwards or downwards) in the relation of the clinically-relevant nucleic acid sequence to the other non-clinically-relevant sequences. Thus, the reference state may be any ratio or other quantity (e.g. other than a 1-1 correspondence), and a measured state signifying a change may be any ratio or other quantity that differs from the reference quantity as determined by the one or more cutoff values.

[0069] The clinically relevant nucleic acid sequence and the background nucleic acid sequence may come from a first type of cells and from one or more second types of cells. For example, fetal nucleic acid sequences originating from fetal/placental cells are present in a biological sample, such as maternal plasma, which contains a background of maternal nucleic acid sequences originating from maternal cells. Thus, in one embodiment, the cutoff value is determined based at least in part on a percentage of the first type of cells in a biological sample. Note the percentage of fetal sequences in a sample may be determined by any fetal-derived loci and not limited to measuring the clinically-relevant nucleic acid sequences. In another embodiment, the cutoff value is determined at least in part on the percentage of tumor sequences in a biological sample, such as plasma, serum, saliva or urine, which contains a background of nucleic acid sequences derived from the non-malignant cells within the body.

[0070] In yet another embodiment, the cutoff value is determined based on an average concentration of a sequence in a plurality of reactions. In one aspect, the cutoff value is determined from a proportion of informative wells that are estimated to contain a particular nucleic acid sequence, where the proportion is determined based on the above-mentioned percentage and/or average concentration. The cutoff value may be determined using many different types of methods, such as SPRT, false discovery, confidence interval, receiver operating characteristic (ROC). This strategy further minimizes the amount of testing required before confident classification can be made. This is of particular relevance to

plasma nucleic acid analysis where the template amount is often limiting. Although presented with respect to digital PCR, other methods may be used.

[0071] Digital PCR involves multiple PCR analyses on extremely dilute nucleic acids such that most positive amplifications reflect the signal from a single template molecule. Digital PCR thereby permits the counting of individual template molecules. The proportion of positive amplifications among the total number of PCRs analyzed allows an estimation of the template concentration in the original or non-diluted sample. This technique has been proposed to allow the detection of a variety of genetic phenomena (Vogelstein, B et al. 1999, supra) and has previously been used for the detection of loss of heterozygosity in tumor samples (Zhou, W. *et al.* 2002, supra) and in the plasma of cancer patients (Chang, HW *et al.* 2002, supra). Since template molecule quantification by digital PCR does not rely on dose-response relationships between reporter dyes and nucleic acid concentrations, its analytical precision should theoretically be superior to that of real-time PCR. Hence, digital PCR could potentially allow the discrimination of finer degrees of quantitative differences between target and reference loci.

[0072] To test this, we first assessed if digital PCR could determine the allelic ratio of *PLAC4* mRNA (Lo, YMD, *et al.* 2007 *Nat Med* 13, 218-223), a placental transcript from chromosome 21, in maternal plasma and thereby distinguish trisomy 21 and euploid fetuses. This approach is referred as the digital RNA-SNP method. We then evaluated whether the increased precision of digital PCR would allow the detection of fetal chromosomal aneuploidies without depending on genetic polymorphisms. We call this digital relative chromosome dosage (RCD) analysis. The former approach is polymorphism-dependent but requires less precision in quantitative discrimination while the latter approach is polymorphism-independent but requires a higher precision for quantitative discrimination.

I. DIGITAL RNA-SNP

A. Overview

[0073] Digital PCR is capable of detecting the presence of allelic ratio skewing of two alleles in a DNA sample. For example, it has been used to detect loss of heterozygosity (LOH) in a tumor DNA sample. Assuming that there are two alleles in the DNA sample, namely A and G, and the A allele would be lost in the cells with LOH. When LOH is present

in 50% of cells in the tumor sample, the allelic ratio of G:A in the DNA sample would be 2:1. However, if LOH is not present in the tumor sample, the allelic ratio of G:A would be 1:1.

[0074] FIG. 1 is a flowchart 100 illustrating a digital PCR experiment. In step 110, the DNA sample is diluted and then distributed to separate wells. Note that the inventors have determined that some plasma nucleic acid species are already quite diluted in the original sample. Accordingly, there is no need for dilution for some templates, if they are already present at the necessary concentrations. In the previous studies (e.g. Zhou et al 2002, supra), a DNA sample is diluted to an extent such that the average concentration of a specific "template DNA" is approximately 0.5 molecule of one of the two templates per well. Note that the term "template DNA" appears to refer to either the A or the G alleles, and that there is no rationale provided for this specific concentration.

[0075] In step 120, in each well, a PCR process is carried out to detect the A and/or the G allele simultaneously. In step 130, the markers in each well are identified (e.g. via fluorescence), e.g. A, G, A and G, or neither. In the absence of LOH, the abundance of the A and the G alleles in the DNA sample would be the same (one copy each per cell). Therefore, the probabilities of a well being positive for the A allele and for the G allele would be the same. This would be reflected by the similar numbers of wells being positive for the A or the G alleles. However, when LOH is present in 50% or greater of cells in a tumor sample, the allelic ratio of the G and the A alleles would be at least 2:1. Previous methods simply assumed that the sample was at least 50% cancerous. Thus, the probability of a well being positive for the G allele would be higher than that for the A allele. As a result, the number of wells being positive for the G allele would be higher than that for the A allele.

[0076] In step 140, to classify the digital PCR results, the number of wells being positive for each allele but not the other would be counted. In the above example, the number of wells being positive for the A allele but negative for the G allele, and the number of wells positive for the G allele but negative for the A allele are counted. In one embodiment, the allele showing less positive wells is regarded as the reference allele.

[0077] In step 150, the total number of informative wells is determined as the sum of the numbers of positive wells for either of the two alleles. In step 160, the proportion (P_r) of informative wells (an example of a parameter) contributed by the allele with more positive wells is calculated.

P_r = No. of wells only positive for the allele with more positive wells / Total no. of wells positive for only one allele (A or G).

Other embodiments could use all wells with one of the alleles divided by all wells with at least one allele.

[0078] In step 170, it is determined whether the value of P_r shows an allelic imbalance. As accuracy and efficiency are desired, this task is not straightforward. One method for determining an imbalance uses a Bayesian-type likelihood method, sequential probability ratio testing (SPRT). SPRT is a method which allows two probabilistic hypotheses to be compared as data accumulate. In other words, it is a statistical method to classify the results of digital PCR as being suggestive of the presence or absence of allelic skewing. It has the advantage of minimizing the number of wells to be analyzed to achieve a given statistical power and accuracy.

[0079] In an exemplary SPRT analysis, the experimental results would be tested against the null and alternative hypotheses. The alternative hypothesis is accepted when there is allelic ratio skewing in the sample. The null hypothesis is accepted when there is no allelic ratio skewing in the sample. The value P_r would be compared with two cutoff values to accept the null or alternative hypotheses. If neither hypothesis is accepted, the sample would be marked as unclassified which means that the observed digital PCR result is not sufficient to classify the sample with the desired statistical confidence.

[0080] The cutoff values for accepting the null or alternative hypotheses have typically been calculated based on a fixed value of P_r under the assumptions stated in the hypotheses. In the null hypothesis, the sample is assumed to exhibit no allelic ratio skewing. Therefore, the probabilities of each well being positive for the A and the G alleles would be the same and, hence, the expected value of P_r would be 1/2. In the alternative hypothesis, the expected value of P_r has been taken to be 2/3 or about halfway between 0.5 and 2/3, e.g. 0.585. Also, due to a limited number of experiments, one can choose an upper bound (.585+3/N) and a lower bound taken as (.585-3/N).

B. Detection of Down Syndrome

[0081] In one embodiment of the present invention, digital SNP is used to detect fetal Down syndrome from a pregnant woman's plasma. Using markers specific to fetal/placental cells, the ratio of alleles in chromosome 21 may be measured. For example, to determine if

an observed degree of overrepresentation of a *PLAC4* allele is statistically significant, SPRT is used.

[0082] According to one exemplary embodiment, digital RNA-SNP determines an imbalance in the ratio of polymorphic alleles of an A/G SNP, rs8130833, located on *PLAC4* mRNA which is transcribed from chromosome 21 and expressed by the placenta. For a heterozygous euploid fetus, the A and G alleles should be equally represented in the fetal genome (1:1 genomic ratio); while in trisomy 21, the trisomic chromosome 21 would be associated with an additional copy of one of the SNP alleles in the fetal genome giving a 2:1 ratio. The aim of the digital PCR analysis is to determine whether the amounts of the two *PLAC4* alleles in the analyzed sample are equal or otherwise. Thus, both the A and G *PLAC4* alleles are the target templates. A real-time PCR assay was designed to amplify *PLAC4* mRNA and the two SNP alleles were discriminated by TaqMan fluorescent probes. A schematic illustration of the analytical steps is shown in Fig. 2A.

[0083] FIG. 2A illustrates a digital RNA-SNP method 200 according to an embodiment of the present invention. In step 210, the sample is received. In step 220, the nucleic acid sequence, e.g. *PLAC4* mRNA, is quantified in extracted RNA samples. In one embodiment, this is done by real-time PCR for *PLAC4* mRNA. In one aspect, this step provides the operator with an idea about how much dilution is required before the target reaches the 'realm' of digital PCR analysis.

[0084] In step 230, the sample is diluted. In step 240, a concentration of the diluted sample is measured. The diluted sample concentration may be confirmed to be ~1 template/well (i.e. reference or non-reference sequence or either allele). Some embodiments use techniques described in section IV for this measurement. For example, we distributed the diluted sample to 96 wells for real-time PCR analysis to confirm that a usable dilution has been achieved. The dilution concentration may also be left as an unknown, thus removing this step as will be explained later.

[0085] In step 250, digital PCR is performed on each well of the array. For example, the same diluted sample was distributed to 384-wells for real-time PCR analysis. From the PCR results, an amount of markers for each nucleic acid sequence and the number of informative wells is identified. An informative well is defined as one that is only positive for the A or G allele but not both. In step 260, a calculation of an expected value of P_r is performed. These steps will be described in more detail later. The calculation includes determining a parameter

from values determined in step 250. For example, the actual average template concentration per well may be calculated.

[0086] In step 270, an SPRT or other likelihood-ratio test may be performed to determine whether or not an imbalance exists. For a euploid case, we expect an equal number of A-positive and G-positive wells. However, when template molecules from a trisomy 21 fetus are analyzed, the number of wells containing just one allele should be higher than that containing just the other allele. In short, allelic imbalance is expected for trisomy 21.

[0087] As mentioned above, SPRT is a Bayesian-type likelihood method which allows two probabilistic hypotheses to be compared as data accumulate. In digital PCR analysis for trisomy 21 detection, the alternative hypothesis is accepted when allelic imbalance exists (i.e. trisomy 21 detected); and the null hypothesis is accepted when there is no allelic imbalance (i.e. trisomy 21 not detected). The allele with the higher number of counts is referred as the potentially overrepresented allele and its proportion among all informative wells (P_r) would be calculated. SPRT is applied to determine if the P_r indicates sufficient degree of allelic imbalance that would be expected for a trisomy 21 sample.

[0088] Operationally, SPRT can be applied and interpreted through the use of graphs with a pair of SPRT curves that are constructed to define the probabilistic boundaries for accepting or rejecting either hypothesis. FIG. 3 illustrates a graph having SPRT curves for determining Down syndrome according to an embodiment of the present invention. The SPRT curves plot the required proportion of informative wells positive for the potentially overrepresented allele P_r, (y-axis) for a given total number of informative wells (x-axis) when confident classification could be made. As depicted in Fig. 3, the upper curve sets the probabilistic boundaries for accepting the alternative hypothesis; while the lower curve sets the probabilistic boundaries for accepting the null hypothesis.

[0089] The experimentally derived P_r value would be compared with the expected value of P_r to accept or reject either hypothesis. If the null hypothesis was accepted, the samples were classified as having been obtained from pregnant women with euploid fetuses. If the alternative hypothesis was accepted, the samples were classified as having been obtained from pregnant women with trisomy 21 fetuses. Alternatively, either hypothesis could not be accepted if the P_r for the given number of informative counts has not yet reached the required level of statistical confidence for disease classification. These cases were deemed unclassifiable until more data were available. If disease classification is not possible,

additional 384-well plates may be performed until the aggregated data become classifiable by SPRT.

[0090] SPRT thus offers the advantage that a smaller amount of testing is required for a given level of confidence than other statistical methods. In practical terms, SPRT allows the acceptance or rejection of either of the hypotheses as soon as the required amount of data has been accumulated and thus minimizes unnecessary additional analyses. This feature is of particular relevance to the analysis of plasma nucleic acids which are generally present at low concentrations where the number of available template molecules is limiting. In addition to a strict classification, the classification may also include a percent accuracy. For example, a classification resulting from a comparison with a cutoff value may provide that a sample shows a likelihood of a nucleic acid sequence imbalance with a certain percentage, or equivalently that determined imbalance is accurate to a certain percentage or other value.

[0091] A similar approach could be applied to determine the genotype of a fetus with regard to either a mutation or genetic polymorphism, using fetal nucleic acids in maternal plasma or serum. One should recall that a fetus would inherit half of its genome from its mother. As an illustration, consider a particular genetic locus with two alleles, A and B. If the mother is a heterozygote with a genotype of AB, the fetus could theoretically have a genotype of AA, BB or AB. If the fetus has a genotype of AB, i.e., the same as that of the mother, then there will only be nucleic acids of the AB genotype (from both the mother and fetus) in maternal plasma. Thus, nucleic acid or allelic balance is seen in maternal plasma. On the other hand, if the fetus has a genotype of AA or BB, then there would be allelic imbalance with an overrepresentation of the A or the B allele, respectively, in maternal plasma. This consideration is also applicable to disease-causing mutations (e.g. those causing cystic fibrosis or beta-thalassemia or spinal muscular atrophy), in which case A could be considered as the wildtype allele and B could be considered as the mutant allele.

II. DIGITAL RCD

[0092] A disadvantage of digital RNA-SNP is that it can only be applied to cases heterozygous for the analyzed SNP. One improvement is that it would be ideal if a noninvasive test for detecting fetal trisomy 21 or other fetal chromosomal aneuploidies (e.g. trisomy 18, 13, and the sex chromosome aneuploidies) based on circulating fetal nucleic acid analysis were independent of the use of genetic polymorphisms. Thus, in one embodiment, chromosome dosage is determined by digital PCR analysis of a non-polymorphic

chromosome 21 locus relative to one located on a reference chromosome, namely chromosome 1 in this study. A change of the ratio of chromosome 21 to chromosome 1 from 2:2 in the genome of a euploid fetus is differentiated from a trisomy 21 case. In digital PCR analysis for trisomy 21 detection, the two hypotheses to be compared would be the null hypothesis that there is no chromosomal imbalance (i.e. trisomy 21 not detected) and the alternative hypothesis that a chromosomal imbalance exists (i.e. trisomy 21 detected).

[0093] This approach can be generalized to the other chromosomes involved in other chromosomal aneuploidies, e.g. chromosome 18 in trisomy 18, chromosome 13 in trisomy 13, chromosome X in Turner syndrome. In addition, apart from chromosome 1, other chromosomes not involved in the aneuploidies concerned can also be used as a reference chromosome. A similar approach can also be applied to the detection of cancer, by analyzing the change of ratio of a chromosome commonly deleted, in part, in cancer, to a reference chromosome. Examples of the former include chromosome 5q in colorectal cancer, chromosome 3p in lung cancer and chromosome 9p in nasopharyngeal carcinoma. FIG. 2B lists some common cancer-related chromosomal aberrations which result in sequence imbalance.

[0094] FIG. 2A also illustrates a digital RCD method 205 according to an embodiment of the present invention. In one embodiment for steps 220-230, extracted DNA is quantified, for example, via Nanodrop techniques and diluted to a concentration of approximately one target template from either chromosomes 21 or the normalizing chromosome (such as chromosome 1) per well. In one embodiment in step 240, the confirmation may be performed by analyzing the diluted DNA sample by the assay using the chromosome 1 probe only in a 96-well format to confirm if ~37% level of the wells were negative before proceeding to digital RCD analysis using both TaqMan probes in 384-well plates. The significance of the 37% will be described later in Section IV.

[0095] The testing of step 240 and results of step 250 may be done with a real-time PCR assay designed to amplify a paralogous sequence (Deutsch, S. et al. 2004 *J Med Genet* 41, 908-915) present on both chromosomes which are distinguished by paralogous sequence variations that are discriminated by a pair of TaqMan probes. In this context, an informative well is defined as one that is positive for either the chromosome 21 or chromosome 1 locus but not both. For a euploid fetus, the number of informative wells positive for either locus should be approximately equal. For a trisomy 21 fetus, there should be an overrepresentation

of wells positive for chromosome 21 than chromosome 1. The exact proportion of the overrepresentation is described in the following sections.

III. INCORPORATING PERCENTAGE OF FETAL SEQUENCES

[0096] A disadvantage of embodiments of methods 200 and 205 described above is that fetal specific markers are required. Accordingly, in one embodiment of the present invention, non-fetal specific markers are used. In order to use such non-fetal specific markers, embodiments of the present invention measure the fractional concentration of fetal DNA in the maternal plasma (i.e. the biological sample). With such information, a more useful value of P_r may be calculated as follows.

[0097] Even with the small fractional percentage of fetal DNA in maternal plasma, a trisomy 21 fetus would contribute an additional dose of chromosome 21 sequences per genome-equivalent (GE) of fetal DNA released into maternal plasma. For example, a maternal plasma sample from a euploid pregnancy containing 50 GE/ml of total DNA with 5 GE/ml DNA contributed by the fetus (i.e. 10% fetal DNA fractional concentration) should contain a total of 100 copies (90 maternal copies + 10 fetal copies) of chromosome 21 sequences per milliliter of maternal plasma. For a trisomy 21 pregnancy, each fetal GE would contribute three copies of chromosome 21, resulting in a total of 105 copies/ml (90 maternal copies + 15 fetal copies) of chromosome 21 sequences in maternal plasma. At 10% fetal DNA concentration, the amount of chromosome 21 derived sequences in the maternal plasma of a trisomic pregnancy would therefore be 1.05 times that of a euploid case. Thus, if an analytical approach could be developed to determine this small degree of quantitative difference, a polymorphism-independent test for noninvasive prenatal diagnosis of fetal trisomy 21 would be achieved.

[0098] Accordingly, the degree of overrepresentation would be dependent on the fractional fetal DNA concentration in the analyzed DNA sample. For example, when placental DNA is analyzed, the theoretical RCD ratio in the fetal genome should be 3:2, i.e. 1.5-fold difference. However, as described above, the theoretical RCD ratio would decrease to 1.05 when a maternal plasma sample containing 10% fetal DNA is analyzed. The experimentally derived P_r is calculated by dividing the number of wells positive only for the chromosome 21 locus by the total number of informative wells. The experimentally derived P_r is subjected to the SPRT analysis with the calculated P_r and the theoretical RCD ratio.

[0099] FIG. 4 shows a method 400 of determining a disease state using a percentage of fetal nucleic acids according to an embodiment of the present invention. In step 410, the fractional percentage of fetal material is measured. In one embodiment, the fractional percentage is determined by measuring the amount of a fetal-specific marker (e.g. Y-chromosome, genetic polymorphism markers (e.g. SNPs), placental epigenetic signatures) in relation to a non-fetal-specific marker (i.e. gene sequence present in both mother and fetus). The actual measurement could be done by real-time PCR, digital PCR, sequencing reactions (including massively parallel genomic sequencing) or any other quantitative methods. In one aspect, it is preferable not to use the gene target that could potentially be in allelic imbalance for this measurement.

[0100] In step 420, digital PCR or other measurement method is performed, including diluting the sample, placing the diluted sample in the wells, and measuring the reactions in each well. In step 430, the PCR results are used to identify markers of different reference nucleic acid sequences (such as chromosomes or alleles). In step 440, the actual proportion (P_r) of the overrepresented sequence is calculated. In step 450, the cutoff value for determining a disease state is calculated using the percentage of fetal material in the sample. In step 460, from the actual P_r and the cutoff value, it is determined whether an imbalance exists.

[0101] In one embodiment, the fractional percentage of reference nucleic acid sequences is incorporated in a digital RNA-SNP method. Thus, when investigating a LOH due to cancer cells, this can be done with tumor samples with less than 50% cancer cells. It also may be used on samples with greater than 50% cancer cells to obtain a more accurate P_r and thus reduce the number of false positives, which would lead to incorrect diagnoses. In another embodiment, the fetal nucleic acid percentage is incorporated in a digital PCR method to determine if a fetus has inherited a parental gene mutation (e.g. that causing cystic fibrosis or beta-thalassemia or spinal muscular atrophy) or polymorphism from maternal plasma nucleic acid analysis.

IV. INCORPORATING AVERAGE CONCENTRATION PER WELL

[0102] Another disadvantage of previous methods (e.g. Zhou, W. *et al.* 2002, supra) is that the average concentration of templates per well (m) is required to be 1 per well. Given that it is difficult to obtain an exact concentration, this can lead to inaccuracies. Furthermore, even with an exact concentration of 1 template per well, previous methods have ignored the

statistical distribution of the templates in a well. In the previous methods, i.e. the old algorithm, the expected value of P_r for accepting the alternative hypothesis is assumed to be the allelic ratio and, thus, is independent of the average concentration of the template DNA per well.

[0103] However, due to a natural statistical variation of the templates in the diluted sample, there will not be exactly 1 template per well. Embodiments of the present invention measure the average concentration of at least one of the sequences, which is then used to calculate the cutoff value, i.e. the expected P_r. In one aspect, this calculation involves a statistical distribution to determine a probability of a well containing the different nucleic acid sequences, which is then used to determine the expected P_r.

[0104] In one embodiment, the average concentration is taken of one reference nucleic acid sequence, which in one instance is nucleic acid sequence with the lower concentration in the DNA sample. In the case of a sample without an imbalance, the concentrations of the two sequences in the sample would be the same and either one can be regarded as the reference allele. In the case of a sample with, for example, LOH, the allele which is deleted in the cancer cells would be regarded as the reference allele. The average concentration of the reference allele would be denoted as m_r. In another embodiment, the sequence with the higher concentration may be taken as the reference sequence.

A. Digital-SNP. Example using SPRT and digital PCR

[0105] FIG. 5 shows a method 500 of determining a disease state using an average template concentration according to an embodiment of the present invention. In step 510, an amount of the different sequences are measured. This may be done for example by counting the makers in a digital PCR experiment as explained above. However, it may be done by other methods that do not involve an amplification step or that do not use a fluorescent marker, but could use other properties, such as physical properties like mass, specific optical properties or base-pairing properties.

[0106] In step 520, the actual proportion of the overrepresented sequence is determined. This may be done as described above by taking the number of wells showing only that sequence and dividing by the number of informative wells. In step 530, the average concentration of at least one of the sequences is measured (the reference sequence). In one embodiment, the reference sequence is the overrepresented sequence. In another

embodiment, the reference sequence is the underrepresented sequence. The measurement may be done by counting the number of wells negative for the reference sequence in the digital PCR experiment. The relationship between the proportion of negative wells and the average template concentration is described by the Poisson distribution as described in the next subsection.

[0107] In step 540, an expected amount of wells positive for the different sequences is calculated, for example, using a Poisson distribution. The expected amount may be as a probability of the sequence per well, an average sequence per well, the number of wells containing the sequence or any other suitable quantity. In step 550, the expected P_r is calculated from the expected amounts. In step 560, a cutoff value is calculated from the expected P_r, for example, by using SPRT. In step 570, a classification of the nucleic acid sequence imbalance is determined. Specific aspects of method 500 are now described.

1. Determining expected amount of sequences

[0108] Once the average concentration per well (reaction or reaction mixture) is known from step 530, the expected number of wells showing that sequence may be calculated in step 540. This amount may be expressed as a %, a fractional value, or an integer value. Using a specific example for illustration, assume the average concentration of the reference template per well (m_r) is 0.5 per well and the genotype of the trisomy 21 fetus at the *PLAC4* SNP, rs8130833, is AGG. Therefore, the reference template would be the A allele and the overrepresented template would be the G allele.

[0109] In one embodiment, a Poisson distribution is assumed for the distribution of the A allele among the reaction mixtures of the wells of the measurement procedure, such as digital PCR. In other embodiments, other distribution functions are used, such as binomial distribution.

[0110] The Poisson equation is: $P(n) = \frac{m^n e^{-m}}{n!}$ where, n=number of template molecules per well; P(n)=probability of n template molecules in a particular well; and m= average number of template molecules in one well in a particular digital PCR experiment.

[0111] Accordingly, the probability of any well not containing any molecule of the A allele at an average A-allele concentration of 0.5 would be:

$$P(0) = \frac{0.5^0 e^{-0.5}}{0!} = e^{-0.5} = 0.6065.$$

[0112] Hence, the probability of any well containing at least one molecule of the A allele would be: 1 – 0.6065 = 0.3935. Therefore, ~39% of the wells would be expected to contain at least one molecule of the A allele.

[0113] As for the non-reference nucleic acid sequence, for each cell of a trisomy 21 fetus, the genomic ratio of A to G would be 1:2. Assuming that the A to G ratio in the extracted RNA or DNA sample would remain unchanged, the average concentration of the G allele per well would be two times that of the A allele, i.e. 2 x 0.5 = 1.

[0114] Accordingly, the probability of any well not containing any molecule of the G allele at an average G-allele concentration of 1 would be:

$$P(0) = \frac{1^0 e^{-1}}{0!} = e^{-1} = 0.3679$$

[0115] Hence, the probability of any well containing at least one molecule of the G allele would be: 1 – 0.3679 = 0.6321. Therefore, ~63% of the wells would be expected to contain at least one molecule of the G allele.

2. Determining proportion of overrepresented sequence

[0116] After the expected amounts are calculated, the proportion of the overrepresented nucleic acid sequence may be determined. Assuming that the filling of the wells with the A allele and the G allele are independent, the probability of a well containing both alleles would be 0.3935 × 0.6321 = 0.2487. Therefore, ~25% of the wells would be expected to contain both alleles.

[0117] The proportion of wells expected to contain the A allele but not the G allele would be the number of wells containing at least one A allele deducted by the number of wells containing both the A and G alleles: 0.3935 – 0.2487 = 0.1448. Similarly, the proportion of wells expected to contain the G allele but not the A allele would be: 0.6321 – 0.2487 = 0.3834. An informative well is defined as a well being positive for either the A allele or the G allele but not both.

[0118] Hence, the expected ratio of wells containing the A allele relative to the G allele in a digital RNA-SNP analysis is 0.1448/0.3834. In other words, the proportion of wells positive

only for the G allele is 2.65 times that of wells positive only for the A allele. This is in contrast to the fetal genomic ratio where the overrepresented allele is 2 times that of the other allele.

[0119] For SPRT analysis, the proportion of the informative wells positive for the overrepresented allele (P_r) is calculated and interpreted using SPRT curves. In the current example, the proportion of informative wells would be: 0.1448 + 0.3834 = 0.5282. Hence, the expected Pr of a trisomy 21 case at m_r 0.5 is: 0.3834/0.5282 = 0.73.

[0120] As the average template concentration (m) is a key parameter in the Poisson equation, the P_r would vary with m. FIG. 6 shows a table 600 that tabulates the expected digital RNA-SNP allelic ratio and P_r of trisomy 21 samples for a range of template concentrations expressed as the average reference template concentration per well (m_r) according to an embodiment of the present invention. Table 600 shows the expected allelic ratio and proportion of informative wells positive for the overrepresented allele (P_r) for a series of average reference template concentrations per well (m_r).

[0121] The expected value of P_r varies with the average concentration of the reference allele per well (m_r) in a non-linear fashion. As shown in table 600, the expected value of P_r for accepting the alternative hypothesis would increase with m_r. As the expected value of P_r for accepting the null hypothesis is fixed at 0.5, the samples with and without allelic imbalance would separate more widely in terms of the value of P_r when m_r increases. Note that in other embodiments the value for accepting the null hypothesis may be other than 0.5. This might occur when the normal ratio is different than 1:1, e.g., 5:3, and thus an imbalance would occur when the ratio deviates from 5:3. The difference in the amounts of the two different nucleic acid sequences would then be determined on a case by case basis.

[0122] However, as previous methods (e.g., Zhou, W. *et al.* 2002, supra) used a fixed expected value of P_r for LOH samples, they underestimated the value of P_r for those samples with LOH (alternative hypothesis accepted). The degree of underestimation would increase with m_r. In other words, the higher the average concentration of the reference allele in the DNA sample, the more inaccurate the old methods would be. This underestimation of P_r for accepting the alternative hypothesis would lead to the inaccurate calculation of the cutoff values for accepting both the null and alternative hypotheses.

3. Calculating cutoff values based on expected P_r

[0123] For embodiments using SPRT, one may use the equations for calculating the upper and lower boundaries of the SPRT curves from El Karoui at al. (2006). Furthermore, the level of statistical confidence preferred for accepting the null or alternative hypothesis could be varied through adjusting the threshold likelihood ratio in the equations. In this study, a threshold likelihood ratio of 8 is used because this value had been shown to provide satisfactory performance to discriminate samples with and without allelic imbalance in the context of cancer detection. Thus, in one embodiment, the equations for calculating the upper and lower boundaries of the SPRT curves are:

Upper boundary = $[(\ln 8)/N - \ln \delta]/\ln \gamma$

Lower boundary = $[(\ln 1/8)/N - \ln \delta]/\ln \gamma$

where, $\delta = (1 - \theta_1)/(1 - \theta_0)$

$\gamma = -(\theta_1(1 - \theta_0) / \theta_0(1 - \theta_1)$

θ_0 = proportion of informative wells containing the non-reference allele if the null hypothesis is true
= 0.5 (see below)

θ_1 = proportion of informative wells containing the non-reference (i.e. overrepresented) allele if the alternative hypothesis is true

N = number of informative wells
= number of wells positive for either allele only

(ln is a mathematical symbol representing the natural logarithm, i.e. $\log_e$.)

[0124] For the determination of θ_0 for accepting the null hypothesis, the sample is assumed to be obtained from a pregnant woman carrying a euploid fetus. Under this assumption, the expected number of wells positive for either template would be 1:1, and thus the expected proportion of informative wells containing the non-reference allele would be 0.5.

[0125] For the determination of θ_1 for accepting the alternative hypothesis, the sample is assumed to be obtained from a pregnant woman carrying a trisomy 21 fetus. The calculations for the expected P_r of trisomy 21 cases for digital RNA-SNP analysis are detailed in Table 600. Hence, θ_1 for digital RNA-SNP analysis refers to the data shown in the last column of table 600.

4. Measurement of average concentration

[0126] The measurement of m_r may be performed through a variety of mechanisms as known or will be known to one skilled in the art. In one embodiment, the value of m_r is determined during the experimental process of digital PCR analysis. As the relationship between the value of m_r and the total number of wells being positive for the reference allele can be governed by a distribution (e.g. the Poisson distribution), m_r can be calculated from the number of wells being positive for the reference allele using this formula:

$m_r = -\ln(1 - \text{proportion of wells being positive for the reference allele})$

Note that ln is the natural logarithm, i.e., $\log_e$. This approach provides a direct and precise estimation of m_r in the DNA sample used for the digital PCR experiment.

[0127] This method may be used to achieve a desired concentration. For example, the extracted nucleic acids of a sample may be diluted to a specific concentration, such as one template molecule per reaction well, as done in step 240 of method 200. In an embodiment using the Poisson distribution, the expected proportion of wells with no template may be calculated as e^{-m}, where m is the average concentration of template molecules per well. For example, at an average concentration of one template molecule per well, the expected proportion of wells with no template molecule is given by e^{-1}, i.e., 0.37 (37%). The remaining 63% of wells will contain one or more template molecules. Typically, the number of positive and informative wells in a digital PCR run would then be counted. The definition of informative wells and the manner by which the digital PCR data are interpreted depends on the application.

[0128] In other embodiments, the average concentration per well, m_r, is measured by another quantification method, for example, quantitative real-time PCR, semi-quantitative competitive PCR, real-competitive PCR using mass spectrometric methods, etc.

B. Digital RCD

[0129] Digital RCD using the average concentration may be performed in a similar manner to the digital SNP method described above. The numbers of wells positive for the reference chromosome (non-chromosome 21) marker, the chromosome 21 marker and both markers can be determined by digital PCR. The average concentration of the reference marker per well (m_r) can be calculated from the total number of wells negative for the reference marker,

irrespective of the positivity of the chromosome 21 marker, according to the Poisson probability function as in the calculation of m_r for the digital SNP analysis.

[0130] SPRT analysis may then be used for classifying a plasma sample as being obtained from a pregnant woman carrying a euploid or a trisomy 21 fetus. The null hypothesis would be accepted when the fetus was euploid. In this scenario, the expected ratio for the wells positive for the reference marker and chromosome 21 marker would be 1:1 and, thus, the expected proportion of informative wells with positive signal for chromosome 21 marker would be 0.5. The alternative hypothesis would be accepted when the fetus was trisomic for chromosome 21. In this scenario, if the sample DNA was solely derived from the fetus, the average concentration of the chromosome 21 marker in each well would be 3/2 times the average concentration of the reference marker (m_r).

[0131] While digital RCD may be used to determine chromosome dosage through the detection of fetal-specific markers, e.g. epigenetic signatures of the placenta (Chim, SSC. *et al.* 2005 *Proc Natl Acad Sci U S A* 102, 14753-14758), an embodiment of the digital RCD analysis uses non-fetal-specific markers. Thus, an additional step of measuring the percentage of fetal material would occur when non-fetal specific markers are used. Therefore, the average concentration of the chromosome 21 marker per well would be dependent on the proportion of the fetal DNA in the sample and can be calculated using: m_r [(200% + fetal DNA percentage)/200%].

[0132] Again using a specific example for illustration, the average concentration of the reference template, chromosome 1, per well (m_r) is assumed to be 0.5 and 50% of the DNA is assumed to be derived from the fetus and 50% of the DNA in the sample is derived from the mother.

[0133] Accordingly, using the Poisson distribution, the probability of any well not containing any molecule of the chromosome 1 locus when its average concentration is 0.5 per well would be:

$$P(0) = \frac{0.5^0 e^{-0.5}}{0!} = e^{-0.5} = 0.6065$$

[0134] Hence, the probability of any well containing at least one molecule of the chromosome 1 locus would be: $1 - 0.6065 = 0.3935$. Therefore, ~39% of the wells would be expected to contain at least one molecule of the locus.

[0135] For each cell of this trisomy 21 fetus, the genomic ratio of chromosome 21 to chromosome 1 would be 3:2. The ratio between chromosome 21 and chromosome 1 in the DNA sample would be dependent on the fractional fetal DNA concentration (fetal DNA %) and would be: 3 × fetal DNA % + 2 (1 – fetal DNA %) : 2 × fetal DNA % + 2 × (1 – fetal DNA %). Thus, in this case when the fractional fetal DNA concentration is 50%, the ratio would be: (3 × 50% + 2 × 50%) / (2 × 50% + 2 × 50%) = 1.25. If the digital SNP method did not use fetal specific markers, such a calculation could also be used to calculate the average concentration of the non-reference sequence.

[0136] Hence, when the average concentration of the chromosome 1 locus per well is 0.5, the average concentration of the chromosome 21 locus per well is: 1.25 × 0.5 = 0.625. Accordingly, the probability of any well not containing any molecule of the chromosome 21 locus when its average concentration is 0.625 per well would be:

$$P(0) = \frac{0.625^0 e^{-0.625}}{0!} = e^{-0.625} = 0.5353$$

[0137] Hence, the probability of any well containing at least one molecule of the chromosome 21 locus would be: 1 – 0.5353 = 0.4647. Therefore, ~46% of the wells would be expected to contain at least one molecule of the locus. Assuming that the filling of the wells with either loci are independent, the probability of a well containing both loci would be 0.3935 × 0.4647 = 0.1829. Therefore, ~18% of the wells would be expected to contain both loci.

[0138] The proportion of wells expected to contain the chromosome 1 locus but not the chromosome 21 locus would be the number of wells containing at least one chromosome 1 locus deducted by the number of wells containing both loci: 0.3935 – 0.1829 = 0.2106. Similarly, the proportion of wells expected to contain the chromosome 21 locus but not both loci would be: 0.4647 – 0.1829 = 0.2818. An informative well is defined as a well positive for either the chromosome 1 locus or the chromosome 21 locus but not both.

[0139] Hence, the expected chromosome 21 to chromosome 1 ratio in digital RCD analysis is 0.2818/0.2106 = 1.34. In other words, the proportion of wells positive only for the chromosome 21 locus is 1.34 times that of wells positive only for the chromosome 1 locus. This is in contrast to the ratio of 1.25 in the DNA sample.

[0140] For SPRT analysis, the proportion of the informative wells positive for the chromosome 21 locus (P_r) would need to be calculated and interpreted using SPRT curves. In the current example, the proportion of informative wells would be: 0.2106 + 0.2818 = 0.4924. Hence, the expected Pr of a trisomy 21 case with 50% fetal DNA at m_r 0.5 is: 0.2818/0.4924 = 0.57.

[0141] As the average template concentration (*m*) is a key parameter in the Poisson equation, the P_r would vary with *m*. FIG. 7 shows a table 700 that tabulates the expected P_r for the fractional fetal DNA concentrations of 10%, 25%, 50% and 100% in trisomy 21 samples at a range of template concentrations expressed as the average reference template concentration per well (m_r) according to an embodiment of the present invention. The calculations for the expected P_r of trisomy 21 cases for digital RCD analyses are detailed in table 700. Hence, θ_1 for digital RCD analysis of samples with varying fetal DNA fractional concentrations can be obtained from the columns showing the corresponding expected P_r-values in table 700.

C. Results

1. Comparison of different m_r

[0142] The basis for the difference between the theoretical (as in the fetal genome) and experimentally-expected degree of allelic or chromosomal imbalance and the calculations to determine the latter for a range of m_r values are shown in tables 600 and 700. In digital RNA-SNP analysis of a trisomy 21 sample, when m_r=0.5, wells containing just the overrepresented allele with respect to wells containing just the reference allele, namely the digital RNA-SNP ratio, is 2.65 (table 600). In digital RCD analysis of a specimen composed of 100% fetal DNA, when m_r=0.5, wells positive solely for the chromosome 21 locus with respect to those positive solely for the chromosome 1 locus, namely the digital RCD ratio, is 1.7 (table 700) (P_r = 0.63, thus the digital RCD ratio is 0.63/(1-0.63) = 1.7). As the fractional fetal DNA concentration decreases, the digital RCD ratio decreases for the same m_r (table 700).

[0143] As shown in tables 600 and 700, the extent of allelic or chromosomal overrepresentation increases with m_r. However, the percentage of informative wells approaches its maximum around m_r=0.5 and decreases gradually with further increase in m_r. In practice, the decline in the proportion of informative wells could be compensated by increasing the total number of wells analyzed if the amount of template molecule for the

specimen is not limiting, but additional wells would require an increase in reagent costs. Hence, optimal digital PCR performance is a trade-off between the template concentration and total number of wells tested per sample.

2. Example using SPRT curves

[0144] As discussed above, the expected degree of allelic or chromosomal imbalance for a digital PCR experiment is dependent on the actual template concentration per reaction mixture (e.g. a well). We describe the template concentration based on the reference allele, i.e. the average reference template concentration per well (m_r). As shown in the above equation, the expected P_r can be used to determine the plotting of the upper and the lower SPRT curves. Since the expected P_r is in turn dependent on the value of m_r, the plotting of the SPRT curves would essentially be dependent on the value of m_r. Thus, in practice, a set of SPRT curves relevant for the actual m_r of a digital PCR dataset would need to be used for the interpretation of the P_r from that particular run.

[0145] FIG. 8 shows a plot 800 illustrating the degree of differences in the SPRT curves for m_r values of 0.1, 0.5 and 1.0 for digital RNA-SNP analysis according to an embodiment of the present invention. Each set of digital PCR data should be interpreted with the specific curves relevant to the exact m_r value of that particular run. Note that since the expected degree of allelic or chromosomal imbalance for the digital RNA-SNP and RCD approaches are different (2:1 for the former and 3:2 for the latter), different sets of SPRT curves are needed for the two digital PCR systems. The experimentally derived P_r is interpreted with the relevant SPRT curves selected by the corresponding m_r of the digital PCR run. This is in contrast to the previous reported use of SPRT for molecular detection of LOH by digital PCR where a fixed set of curves had been used.

[0146] The practical manner for interpreting the digital PCR data using SPRT is illustrated below using a hypothetical digital RNA-SNP run. After digital RNA-SNP analysis of each case, the number of wells positive for the A allele only, the G allele only or both alleles are counted. The reference allele is defined as the allele with the smaller number of positive wells. The value of m_r is calculated using the total number of wells negative for the reference allele, irrespective of whether the other allele is positive, according to the Poisson probability density function. The data of our hypothetical example are as follows:

[0147] In a 96-well reaction, 20 wells are positive for the A allele only, 24 wells are positive for the G allele only, and 33 wells are positive for both alleles. The A allele is regarded as the reference allele because there are less A-positive than G-positive wells. The number of wells negative for the reference allele is $96 - 20 - 33 = 43$. Therefore, m_r can be calculated using the Poisson equation: $-\ln(43/96) = 0.80$. The experimentally determined P_r of this case is: $24 / (20 + 24) = 0.55$.

[0148] According to table 600, the expected P_r of a trisomy 21 sample at $m_r = 0.8$ is 0.76. Thus, θ_1 is 0.76 for this case. The SPRT curves based on $\theta_1 = 0.76$ would be used to interpret the experimentally derived P_r of this case which is 0.55. When P_r=0.55 is fitted onto the relevant SPRT curves, the data point falls under the lower curve. Hence, the case is classified as euploid, see FIG. 3.

3. Comparison to old method

[0149] FIG. 9A shows a table 900 of a comparison of the effectiveness of the new and old SPRT algorithms for classifying euploid and trisomy 21 cases in 96-well digital RNA-SNP analyses. FIG. 9B shows a table 950 of a comparison of the effectiveness of the new and old SPRT algorithms for classifying euploid and trisomy 21 cases in 384-well digital RNA-SNP analyses. The new algorithm refers to the selection of SPRT curves specific for the m_r derived from the digital PCR data. The old algorithm refers to the use of a fixed set of SPRT curves for all digital PCR runs. The effect of incorrect calculation of the cutoff values on the accuracy of classification is revealed by the simulation analysis shown in table 900.

[0150] Compared with the use of a fixed set of SPRT curves in previous studies, the proportion of unclassifiable data is much lower with our approach, as shown in tables 900 and 950. For example, using our approach, at m_r=0.5, 14% and 0% of trisomy 21 samples would be unclassifiable for 96-well and 384-well digital RNA-SNP analyses, respectively, but 62% and 10%, respectively, with the use of fixed curves (Table 900). Hence, our approach allows disease classification with lesser number of informative wells.

[0151] As shown in table 900, the new algorithm is more accurate in classifying the samples as having or not having allelic ratio skewing for all values of m_r from 0.1 to 2.0. For example, when m_r equals 1.0 and a 96–well digital RNA-SNP run is performed, the new algorithm correctly classifies 88% and 92% of samples with and without allelic ratio

skewing, respectively, whereas the percentage of correct classification of samples with and without allelic ratio skewing using the old algorithm is only 19% and 36%, respectively.

[0152] Using the new algorithm, the separation of samples with and without allelic ratio skewing would increase with m_r. As a result, the classification accuracies would increase with m_r. The effect of increase in separation of the two groups of samples on the classification accuracy would diminish when m_r increases to beyond 2.0 because of the reduction in the percentage of informative wells. In contrast, using the old algorithm, the classification accuracies significantly reduce when m_r increases because of the increased deviation of expected P value from its true value.

[0153] Our experimental and simulation data show that digital RNA-SNP is an effective and accurate method for trisomy 21 detection. As *PLAC4* mRNA in maternal plasma is derived purely from the fetus, for 12 of the 13 maternal plasma samples tested, only one 384-well digital PCR experiment was required for correct classification (Table 1350 of FIG. 13B). This homogenous, real-time digital PCR-based approach thus offers an alternative to the mass spectrometry-based approach for RNA-SNP analysis (Lo, YMD, *et al.* 2007 *Nat Med*, supra). Apart from placental-specific mRNA transcripts, we also envision that other types of fetal-specific nucleic acid species in maternal plasma could be used for digital PCR-based detection of fetal chromosomal aneuploidies. One example is fetal epigenetic markers (Chim, SSC et al. (2005) *Proc Natl Acad Sci U S A* 102, 14753-14758; Chan, KCA *et al.* (2006) *Clin Chem* 52, 2211-2218), which have recently been used for the noninvasive prenatal detection of trisomy 18 using the epigenetic allelic ratio (EAR) approach (Tong, YK et al. (2006) *Clin Chem* 52, 2194-2202). Thus, we predict that digital EAR would be a possible analytical technique.

V. INCREASING %, MULTIPLE MARKERS, AND PCR ALTERNATIVES

[0154] As described above, the application of embodiments of the present invention to DNA extracted from maternal plasma can be complicated when the fetal DNA only constitutes a minor fraction of maternal plasma DNA, with a mean fractional concentration of some 3% between weeks 11 and 17 of gestation. Nevertheless, as shown herein, digital RCD allows aneuploidy detection even when the aneuploid DNA is present as a minor population. With a decreasing fractional concentration of fetal DNA, such as may be present during early gestation, a larger number of informative counts is needed for digital RCD. A significance of the present work, as summarized in table 1200 of FIG. 12, is that we have provided a set of

benchmark parameters, e.g. fractional fetal DNA and total template molecules required, which diagnostic assays can be built upon. In our opinion, 7680 total number of reactions for a fractional fetal DNA concentration of 25% is a particularly attractive set of benchmark parameters. These parameters should allow euploid and trisomy 21 samples to be classifiable correctly 97% of the time, as shown in table 1200.

[0155] The number of plasma DNA molecules that are present per unit volume of maternal plasma is limited (Lo, YMD. *et al.* 1998 *Am J Hum Genet* 62, 768-7758). For example, in early pregnancy, the median maternal plasma concentration of an autosomal locus, the *β-globin* gene, has been shown to be 986 copies/mL, with contributions from both the fetus and mother (Lo, YMD. *et al.* 1998 *Am J Hum Genet* 62, 768-7758). To capture 7,680 molecules, DNA extracted from some 8 mL of maternal plasma would be needed. This volume of plasma, obtainable from ~15 mL of maternal blood, is at the limit of routine practice. However, we envision that multiple sets of chr21 and reference chromosome targets can be combined for digital RCD analysis. For 5 pairs of chr21 and reference chromosome targets, just 1.6 mL of maternal plasma would be needed to provide the number of template molecules needed for analysis. Multiplex single molecule PCR could be performed. The robustness of such multiplex single molecule analysis has been demonstrated previously for single molecule haplotyping (Ding, C. and Cantor, CR. 2003 *Proc Natl Acad Sci U S A* 100, 7449-7453).

[0156] Alternatively, to achieve a fractional fetal DNA concentration of 25%, methods may allow the selective enrichment of fetal DNA (Li, Y. et al. 2004 *Clin Chem* 50, 1002-1011) or the suppression of the maternal DNA background (Dhallan, R et al. 2004 *JAMA* 291, 1114-1119) or both, in maternal plasma. Apart from such physical methods for fetal DNA enrichment and maternal DNA suppression, it would also be possible to use a molecular enrichment strategy, such as by targeting fetal DNA molecules which exhibit a particular DNA methylation pattern (Chim, SSC et al, 2005 *Proc Natl Acad Sci USA* 102,14753-14758, Chan, KCA et al. 2006 *Clin Chem* 52, 2211-2218; Chiu, RWK *et al.* 2007 *Am J Pathol* 170, 941-950.)

[0157] Additionally, there are now a number of alternative approaches to the manual set up of digital real-time PCR analyses as used in the current study for conducting digital PCR. These alternative approaches include microfluidics digital PCR chips (Warren, L et al. 2006 *Proc Natl Acad Sci U S A* 103, 17807-17812; Ottesen, EA et al. 2006 *Science* 314, 1464-

1467), emulsion PCR (Dressman, D et al. 2003 *Proc Natl Acad Sci U S A* 100, 8817-8822), and massively parallel genomic sequencing (Margulies, M. *et al.* 2005 *Nature* 437, 376-380) using for example the Roche 454 platform, the Illumina Solexa platform, and the SOLiD™ system of Applied Biosystems, etc. With regard to the latter, our method is also applicable to massively parallel sequencing methods on single DNA molecules, which do not require an amplification step, such as the Helicos True Single Molecule DNA sequencing technology (Harris TD et al. 2008 Science, 320, 106-109), the single molecule, real-time (SMRT™) technology of Pacific Biosciences, and nanopore sequencing (Soni GV and Meller A. 2007 *Clin Chem* 53, 1996-2001). With the use of these methods, digital RNA-SNP and digital RCD could be performed rapidly on a large number of samples, thus enhancing the clinical feasibility of the methods proposed here for noninvasive prenatal diagnosis.

EXAMPLES

The following examples are offered to illustrate, but not to limit the claimed invention.

I. COMPUTER SIMULATIONS

[0158] Computer simulation was performed to estimate the accuracy of diagnosing trisomy 21 using the SPRT approach. The computer simulation was performed with the Microsoft Excel 2003 software (Microsoft Corp., USA) and SAS 9.1 for Windows software (SAS Institute Inc., NC, USA). The performance of digital PCR is an interplay between the reference template concentration (m_r), number of informative counts and projected degree of allelic or chromosomal imbalance (P_r). Separate simulations were performed for a range of each of these variables. Since the decision boundaries of the SPRT curves for digital RNA-SNP and digital RCD were different, the simulation analyses for the two systems were performed separately.

[0159] For each digital PCR condition simulated (i.e. m_r, fetal DNA fractional concentration, total well number), two rounds of simulation were performed. The first round simulated the scenario that the tested samples were obtained from pregnant women carrying euploid fetuses. The second round simulated the scenario when the tested samples were obtained from pregnant women carrying trisomy 21 fetuses. For each round, testing of 5000 fetuses was simulated.

A. RNA-SNP

[0160] For digital RNA-SNP, simulations of a 384-well experiment with m_r=0.1 to m_r=2.0 were performed. At each m_r value, we simulated the scenario whereby 5000 euploid and 5000 trisomy 21 fetuses were tested. The SPRT curves appropriate for the given m_r were used to classify the 10,000 fetuses. FIG. 10 is a table 1000 showing the percentages of fetuses correctly and incorrectly classified as euploid or aneuploid and those not classifiable for the given informative counts according to an embodiment of the present invention. The accuracies for diagnosing euploid and aneuploid cases are both 100%, for m_r between 0.5 and 2.0. When m_r=0.1, only 57% and 88% of euploid and trisomy 21 fetuses could be accurately classified after the analysis of 384 wells.

[0161] The simulation data were generated as described in the following steps:

[0162] In step **1**, for each well, two random numbers were generated using the Random(Poisson) function of the SAS program (www.sas.com/technologies/analytics/statistics/index.html) to represent the A and the G alleles, respectively. The Random(Poisson) function would generate positive integers starting from 0 (i.e. 0, 1, 2, 3,....) and the probability of each integer being generated was dependent on the probability of this number according to the Poisson probability density function for a given mean value which represented the average concentration of the alleles per well. A well was regarded as positive for the A allele if the random number representing the A allele was larger than zero, i.e. contained 1 or more molecules of the A allele. Similarly, the well was regarded as positive for the G allele if the random number representing the G allele was larger than zero.

[0163] To simulate the scenario of a pregnant woman carrying a euploid fetus, the same mean value was used for generating the random numbers for the A allele and the G allele. For example, in the analysis simulating digital RNA-SNP analyses at m_r = 0.5, the mean value for either the A allele of the G allele was set identically to 0.5 which meant an average concentration for either allele of 0.5 molecule per well. Using the Poisson equation, at a mean concentration of 0.5, the proportion of wells being positive for the A or the G alleles would be the same and was 0.3935, see table 600.

[0164] When simulating the digital RNA-SNP analysis of a pregnant woman with a trisomy 21 fetus at m_r = 0.5, the average concentration of the overrepresented allele per well

would be expected to be 2 times of that of the reference allele, i.e. 1. In this situation, the probability of a well being positive for the overrepresented allele was 0.6321, see table 600.

[0165] After generating a random number for a digital PCR well, the well could be classified as one of the following statuses:

a. negative for both the A and the G alleles
b. positive for both the A and the G alleles
c. positive for the A allele but negative for the G allele
d. positive for the G allele but negative for the A allele

[0166] In step 2, step 1 was repeated until the desired number of wells, 384 wells for the current simulation, had been generated. The numbers of wells positive for the A allele only and the G allele only were counted. The allele with less positive wells was regarded as the reference allele and the allele with more positive wells was regarded as the potentially overrepresented allele. The number of the informative wells was the total number of wells positive for either allele but not both. The proportion of informative wells containing the potentially overrepresented allele (P_r) was then calculated. The upper and lower boundaries for the relevant SPRT curves to accept the null or alternative hypothesis were calculated according to an embodiment of the present invention.

[0167] In step 3, 5000 simulations were performed for each of the two scenarios of the pregnant woman carrying a euploid or a trisomy 21 fetus. Each simulation could be regarded as an independent biological sample obtained from pregnant women. In Table 1000, the correct classification of euploid cases refers to those euploid cases in which the null hypothesis was accepted and the incorrect classification of euploid cases refers to those euploid cases in which the alternative hypothesis was accepted. Similarly, those trisomy 21 cases in which the alternative hypothesis was accepted were regarded as correctly classified and those trisomy 21 cases in which the null hypothesis was accepted were regarded as incorrectly classified. For both groups, those cases in which neither the null or alternative hypothesis was accepted after the pre-specified total number of wells had been simulated were regarded as unclassified.

[0168] In step 4, steps 1 to 3 were performed for m_r ranging from 0.1 to 2.0 at increments of 0.1.

B. RCD

[0169] FIG. 11 is a table 1100 showing computer simulations for digital RCD analysis for a pure (100%) fetal DNA sample for m_r ranging from 0.1 to 2.0 according to an embodiment of the present invention. As the fractional fetal DNA concentration becomes lower, the degree of chromosome 21 overrepresentation diminishes and thus a larger number of informative wells for accurate disease classification is required. Hence, simulations were further performed for fetal DNA concentrations of 50%, 25% and 10% for a total well number ranging from 384 to 7680 wells at m_r=0.5.

[0170] FIG. 12 is a table 1200 showing results of computer simulation of accuracies of digital RCD analysis at m_r=0.5 for the classification of samples from euploid or trisomy 21 fetuses with different fractional concentrations of fetal DNA according to an embodiment of the present invention. The performance of digital RCD is better for cases with a higher fetal DNA fractional concentration. At 25% fetal DNA concentration and a total number of 7680 PCR analyses, 97% of both euploid and aneuploid cases would be classifiable with no incorrect classification. The remaining 3% of cases require further analyses until classification can be achieved.

[0171] The procedures for simulating digital RCD analyses were similar to those described for digital RNA-SNP analysis. The steps for the simulations are described below:

[0172] In step 1, two random numbers under the Poisson probability density function were generated to represent the reference locus, chromosome 1, and the chromosome 21 locus. For subjects carrying euploid fetuses, the average concentrations of both the chromosome 1 and chromosome 21 loci were the same. In this simulation analysis, an average template concentration of 0.5 for each locus per well was used. For subjects carrying trisomy 21 fetuses, the m_r in this simulation was 0.5 but the average concentration of the chromosome 21 locus per well would depend on the fractional fetal DNA concentration in the tested sample, as shown in Table 700. The distribution of the reference and/or the chromosome 21 loci to a well was determined by the random numbers representing the respective locus which were generated according to the Poisson probability density function with the appropriate average concentration of the locus per well.

[0173] In step 2, step 1 was repeated until the desired number of wells had been generated, e.g. 384 wells for a 384-well plate experiment. The numbers of wells positive for chromosome 1 only and chromosome 21 only were counted. The number of the informative

wells was the total number of wells positive for either one of the chromosomes but not both. The proportion of informative wells positive for chromosome 21 (P_r) was then calculated. The upper and lower boundaries of the relevant SPRT curves to accept the null or alternative hypothesis were calculated as described in the earlier section on SPRT analysis.

[0174] In step 3, 5000 simulations were performed for each of the two scenarios of the pregnant woman carrying a euploid or a trisomy 21 fetus. Each simulation could be regarded as an independent biological sample obtained from pregnant women. In Table 1100, the correct classification of euploid cases refers to those euploid cases in which the null hypothesis was accepted and the incorrect classification of euploid cases refers to those euploid cases in which the alternative hypothesis was accepted. Similarly, those trisomy 21 cases in which the alternative hypothesis was accepted were regarded as correctly classified and those trisomy 21 cases in which the null hypothesis was accepted were regarded as incorrectly classified. For both groups, those cases in which neither the null or alternative hypothesis was accepted after the pre-specified total number of wells had been simulated were regarded as unclassified.

[0175] In step 4, steps 1 to 3 were repeated for samples with 10% 25%, 50% and 100% fetal DNA at total well numbers ranging from 384 to 7680.

II. VALIDATION OF TRISOMY 21 DETECTION

A. RNA-SNP for *PLAC4*

[0176] The practical feasibility of digital RNA-SNP was demonstrated using the rs8130833 SNP on the *PLAC4* gene on chromosome 21 (Lo, YMD et al. 2007 *Nat Med* 13, 218-223). Placental DNA and RNA samples from two euploid and two trisomy 21 heterozygous placentas were analyzed. The placental DNA samples were analyzed with the digital RNA-SNP protocol but with the omission of the reverse transcription step, thus essentially converting the procedure to digital *DNA*-SNP analysis. To strike a balance between the chance of correct case classification and the proportion of informative wells, we diluted the samples aiming for one allele of any type per well and confirmed by a 96-well digital PCR analysis. This was followed by a 384-well digital RNA-SNP experiment. P_r and m_r were calculated and the SPRT curve for this m_r value was used for data interpretation.

[0177] FIG. 13A shows a table 1300 of digital RNA-SNP analysis in placental tissues of euploid and trisomy 21 pregnancies according to an embodiment of the present invention.

Genotypes were determined by mass spectrometric assay. "Euploid" was assigned when the experimentally obtained P_r was below the unclassifiable region; "T21", representing trisomy 21, was assigned when the experimentally obtained P_r was above the unclassifiable region. T21, trisomy 21. Each of these cases was correctly classified, using both the DNA and RNA samples, with one 384-well experiment.

[0178] We further tested plasma RNA samples from nine women carrying euploid fetuses and four women carrying trisomy 21 fetuses. FIG. 13B shows a table 1350 of digital RNA-SNP analysis of maternal plasma from euploid and trisomy 21 pregnancies according to an embodiment of the present invention. All of the cases were correctly classified. Initial results for one trisomy 21 case (M2272P) fell within the unclassifiable region between the SPRT curves after one 384-well experiment. Thus, an additional 384-well run was performed. New m_r and P_r values were calculated from the aggregated data from the total of 768 wells and the classification was performed using a new set of SPRT curves selected based on this m_r value. The case was then scored correctly as aneuploid.

[0179] Our experimental and simulation data show that digital RNA-SNP is an effective and accurate method for trisomy 21 detection. As *PLAC4* mRNA in maternal plasma is derived purely from the fetus, for 12 of the 13 maternal plasma samples tested, only one 384-well digital PCR experiment was required for correct classification. This homogenous, real-time digital PCR-based approach thus offers an alternative to the mass spectrometry-based approach for RNA-SNP analysis. Apart from placental-specific mRNA transcripts, we also envision that other types of fetal-specific nucleic acid species in maternal plasma could be used for digital PCR-based detection of fetal chromosomal aneuploidies. One example is fetal epigenetic markers which have recently been used for the noninvasive prenatal detection of trisomy 18 using the epigenetic allelic ratio (EAR) approach (Tong YK et al. 2006 *Clin Chem,* 52, 2194-2202). Thus, we predict that digital EAR would be a possible analytical technique.

B. RCD

[0180] The practical feasibility of digital RCD for trisomy 21 detection was also investigated using a PCR assay targeting paralogous sequences on chromosome 21 and 1. Paralogous loci were used here by way of examples. Non-paralogous sequences on chromosome 21 and any other reference chromosome can also be used for RCD. Placental DNA samples from two euploid and two trisomy 21 placentas were diluted to approximately

one target template from either chromosome per well and confirmed by a 96-well digital PCR analysis. Each confirmed sample was analyzed by a 384-well digital RCD experiment and the values of P_r and m_r were calculated. For digital RCD, the chromosome 1 paralog was the reference template. This m_r value was used to select a corresponding set of SPRT curves for data interpretation. All of the placental samples were correctly classified as shown in FIG. 14A.

[0181] To demonstrate that the digital RCD approach can be used to detect trisomy 21 DNA which is mixed with an excess of euploid DNA, such as the scenario of fetal DNA in maternal plasma, mixtures containing 50% and 25% of trisomy 21 placental DNA in a background of euploid maternal blood cell DNA were analyzed. Placental DNA from 10 trisomy 21 and 10 euploid cases was each mixed with an equal amount of euploid maternal blood cell DNA, thus producing twenty 50% DNA mixtures. FIG. 14B shows a plot 1440 illustrating the SPRT interpretation for RCD analysis of the 50% fetal DNA mixtures according to an embodiment of the present invention. Similarly, placental DNA from 5 trisomy 21 and 5 euploid cases was each mixed with 3 times excess of euploid maternal blood cell DNA, thus producing ten 25% DNA mixtures. FIG. 14C shows a plot 1470 illustrating the SPRT interpretation for RCD analysis of the 25% fetal DNA mixtures. All the euploid and aneuploid DNA mixtures were correctly classified, as shown in FIGS. 14B and 14C.

[0182] Each sample reached the point of being classifiable after a number of 384-well digital PCR analyses as marked on FIGS. 14B and 14C. For the 50% DNA mixtures, the number of 384-well plates required ranged from one to five. For the 25% DNA mixtures, the number of 384-well plates required ranged form one to seven. The cumulative proportion of cases being correctly classified with the addition of every 384 digital PCR analyses were as predicted by the computer simulation presented in Table 1200.

III. METHOD WITH DIGITAL PCR

A. Digital RNA-SNP

[0183] All RNA samples were first reverse transcribed with a gene-specific reverse transcription primer using the ThermoScript reverse transcriptase (Invitrogen). Sequence of the reverse transcription primer was 5′-AGTATATAGAACCATGTTTAGGCCAGA-3′ (Integrated DNA Technologies, Coralville, IA). The subsequent treatment of the reverse transcribed RNA (i.e. the cDNA) samples for digital RNA-SNP, and DNA samples (e.g.

placental DNA) was essentially the same. Prior to digital PCR analysis, DNA and the cDNA samples were first quantified using a real-time PCR assay towards *PLAC4*, consisting of primers 5′-CCGCTAGGGTGTCTTTTAAGC-3′, 5′-GTGTTGCAATACAAAATGAGTTTCT-3′, and the fluorescent probe 5′-(FAM)ATTGGAGCAAATTC(MGBNFQ)-3′ (Applied Biosystems, Foster City, CA), where FAM is 6-carboxyfluorescein and MGBNFQ is a minor groove binding non-fluorescent quencher.

[0184] A calibration curve was prepared by serial dilutions of HPLC-purified single-stranded synthetic DNA oligonucleotides (Proligo, Singapore) specifying the amplicon. The sequence was 5′-CGCCGCTAGGGTGTCTTTTAAGCTATTGGAGCAAATTCAAATTTGGCTTAAAGAAAAAGAAACTCATTTTGTATTGCAACACCAGGAGTATCCCAAGGGACTCG-3′. The reaction was set up using 2X TaqMan Universal PCR Master Mix (Applied Biosystems) in a reaction volume of 25 μL. 400 nM of each primer and 80 nM of the probe were used in each reaction. The reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 45 cycles of 95°C for 15s and 60°C for 1 min in an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems). Serial dilutions of the DNA or cDNA samples were then undertaken such that the subsequent digital PCR amplification could be performed at approximately one template molecule per well. At such a concentration, it was expected that approximately 37% of the reaction wells would show negative amplification and was first confirmed by conducting a 96-well digital real-time PCR analysis. This was followed by digital RNA-SNP analysis conducted in 384-well plates using a set of non-intron spanning primers: the forward primer 5′-TTTGTATTGCAACACCATTTGG-3′ and the gene-specific reverse transcription primer described above.

[0185] Two allele-specific TaqMan probes targeting each of the two alleles of the rs8130833 SNP on the *PLAC4* sequence were designed. Their sequences were 5′-(FAM)TCGTCGTCTAACTTG(MGBNFQ)-3′ and 5′-(VIC)ATTCGTCATCTAACTTG(MGBNFQ) for the G and A alleles, respectively. The reaction was set up using 2X TaqMan Universal PCR Master Mix in a reaction volume of 5 μL. Each reaction contains 1X TaqMan Universal PCR Master Mix, 572 nM of each primer, 107 nM of the allele-G-specific probe and 357 nM of the allele-A-specific probe. The reaction was carried out in the ABI PRISM 7900HT Sequence Detection System. The reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 45 cycles of 95°C

for 15s and 57°C for 1 min. During the reaction, the fluorescence data were collected by the "Absolute Quantification" application of the SDS 2.2.2 software (Applied Biosystems). The software automatically calculated the baselines and the threshold values. The number of wells which were positive for either the A or the G alleles was recorded and subjected to SPRT analysis.

B. Digital RCD analysis

[0186] All placental and maternal buffy coat DNA samples used in this study were first quantified by the NanoDrop spectrophotometer (NanoDrop Technology, Wilmington, DE). The DNA concentration is converted to copies/μL using a conversion of 6.6 pg/cell. The amount of DNA corresponding to approximately one template per well was determined by serially diluting the DNA samples and confirmed with the real-time PCR assay in a 96-well format where we expect approximately 37% of the wells to show negative amplification. The PCR setup for the confirmatory plate was the same as described below except that only the probe for the reference chromosome was added. In the digital RCD analysis, the paralogous loci on chromosome 21 and 1 (Deutsch, S. *et al.* 2004 *J Med Genet* 41, 908-915) were first co-amplified by forward primer 5'-GTTGTTCTGCAAAAAACCTTCGA-3' and reverse primer 5'-CTTGGCCAGAAATACTTCATTACCATAT-3'. Two chromosome-specific TaqMan probes were designed to target the chromosome 21 and 1 paralogs, and their sequences were 5'-(FAM)TACCTCCAT AATGAGTAAA(MGBNFQ)-3' and 5'-(VIC)CGTACCTCTGTAATGTGTAA(MGBNFQ)-3', respectively. Each reaction contained 1X TaqMan Universal PCR Master Mix (Applied Biosystems), 450 nM of each primer, and 125 nM of each probe. The total reaction volume was 5μL/well. The reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 50 cycles of 95°C for 15s and 60°C for 1 min. All real-time PCR experiments were carried out on an ABI PRISM 7900HT Sequence Detection System (Applied Biosystems), and the fluorescence data were collected by the "Absolute Quantification" application of the SDS 2.2.2 software (Applied Biosystems). The default baselines and manual threshold values were used. The number of wells which were positive for either chromosome 21 or chromosome 1 was recorded and subjected to SPRT analysis. One or more 384-well plates would be analyzed until disease classification was possible by SPRT.

IV. USING MICROFLUIDICS-BASED DIGITAL PCR

A. Digital RNA-SNP

[0187] This example demonstrates the performance of digital PCR analysis using microfluidics-based digital PCR. A variant of this approach is illustrated here, by way of example but not by way of limitation to, using a Fluidigm BioMark™ System. This system can perform over 9000 digital PCRs per run.

[0188] Placental tissues and maternal peripheral blood samples were obtained from pregnant women carrying euploid or trisomy 21 fetuses. Genotyping of the rs8130833 SNP on the *PLAC4* gene was carried out in placental DNA samples by primer extension followed by mass spectrometry. RNA was extracted from the placental and maternal plasma samples.

[0189] All RNA samples were reverse transcribed with a gene-specific reverse transcription primer (5′-AGTATATAGAACCATGTTTAGGCCAGA-3′) using the ThermoScript reverse transcriptase (Invitrogen). For the placental cDNA samples, serial dilutions were performed such that the subsequent digital PCR amplification could be performed at approximately one template molecule per well.

[0190] Digital PCR was conducted on the BioMark System™ (Fluidigm) with a 12.765 Digital Array (Fluidigm). Each Digital Array consists of 12 panels for accommodating 12 sample-assay mixtures. Each panel is further partitioned into 765 wells for carrying out a 7-nL reaction/well. The rs8130833 SNP region on the *PLAC4* gene was amplified by the forward primer (5′-TTTGTATTGCAACACCATTTGG-3′) and the gene-specific reverse transcription primer described above. Two allele-specific TaqMan probes targeting each of the two alleles of the rs8130833 SNP were designed. Their sequences were 5′-(FAM)TCGTCGTCTAACTTG(MGBNFQ)-3′ and 5′-(VIC)ATTCGTCATCTAACTTG(MGBNFQ) for the G and A alleles, respectively. The reaction for one array panel was set up using 2X TaqMan Universal PCR Master Mix in a reaction volume of 10 μL. Each reaction contains 1X TaqMan Universal PCR Master Mix, 572 nM of each primer, 53.5 nM of the allele-G-specific probe, 178.5 nM of the allele-A-specific probe and 3.5 μL of the cDNA sample. One reaction panel was used for each placental cDNA sample while 12 panels were used for each maternal plasma sample. The sample-assay mixtures were loaded into the Digital Array by a NanoFlex™ IFC controller (Fluidigm). The reaction was carried out in the BioMark™ System. The reaction was

initiated at 50°C for 2 min, followed by 95°C for 10 min and 40 cycles of 95°C for 15s and 57°C for 1 min.

[0191] Placental RNA samples from one euploid and two T21 heterozygous placentas were analyzed in a 765-well reaction panel. For each sample, the number of informative wells, comprising the ones positive for either the A or the G allele (but not both), was counted. The proportion of the overrepresented allele among all the informative wells (Pr) was determined. SPRT curves appropriate for the exact average reference template concentration per well (m_r) of these runs were applied to determine if the experimentally-obtained Pr indicated a euploid or T21 sample. As shown in FIG. 15A, all RNA samples were correctly classified using this approach.

[0192] We further tested the plasma RNA samples from four women carrying euploid fetuses and one woman carrying a trisomy 21 fetus. Each sample was analyzed in twelve 765-well reaction panels, i.e. 9180 reactions per plasma RNA sample. FIG. 15B shows the number of informative wells for each of the 12 panels for this plasma RNA sample. As shown in the table, the template concentration in the plasma sample was so diluted that the number of informative wells in any one reaction panel was not sufficient for the SPRT classification. The informative wells from three reaction panels had to be combined before this sample was classified as a euploid sample (FIG. 15C). FIG. 15C shows that using the aggregated data from two to twelve panels, all of the plasma cases could be correctly classified.

[0193] Compared with the manual method for performing digital PCR, this microfluidics-based method is much more rapid and less labor-intensive. The whole process could be completed in two and a half hours.

B. Digital RNA-SNP analysis for the prenatal detection of trisomy 18

[0194] In this example, we used a digital PCR-based allelic discrimination assay on *serpin peptidase inhibitor clade B (ovalbumin) member 2* (*SERPINB2*) mRNA, a placenta-expressed transcript on chromosome 18, to detect an imbalance in the ratio of polymorphic alleles for trisomy 18 fetuses. Extraction of DNA and RNA from placental tissue samples was performed using the QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) and the TRIzol reagent (Invitrogen, Carlsbad, CA), respectively, as described in the manufacturers' protocols. The extracted placental RNA samples were subjected to DNase I treatment (Invitrogen) for removal of contaminating genomic DNA. Genotyping of the rs6098 SNP on

the SERPINB2 gene was carried out in placental tissue DNA samples with a Homogenous MassEXTEND (hME) assay using the MassARRAY Compact (Sequenom, San Diego) as previously described.

[0195] Reverse transcription for the SERPINB2 transcript was performed on the placental tissue RNA samples with a gene-specific primer 5′-CGCAGACTTCTCACCAAACA-3′ (Integrated DNA Technologies, Coralville, IA) using the ThermoScript reverse transcriptase (Invitrogen). All cDNA samples were diluted to a concentration such that the subsequent digital PCR amplification could be performed at an average concentration of one template molecule per reaction well. Digital PCR was set up using the TaqMan Universal PCR Master Mix (Applied Biosystems, Foster City, CA) and the BiomarkTM PCR Reagents (Fluidigm, San Francisco). The forward primer 5′- CTCAGCTCTGCAATCAATGC-3′ (Integrated DNA Technologies) and the reverse primer (identical to the gene-specific primer used for reverse transcription) were used at a concentration of 600 nM. The two TaqMan probes targeting the A or G allele of the rs6098 SNP on the SERPINB2 sequence were 5′-(FAM)CCACAGGGAATTATTT (MGBNFQ)-3′ and 5′-(FAM)CCACAGGGGATTATTT(MGBNFQ)-3′ (Applied Biosystems). FAM is 6-carboxyfluorescein and MGBNFQ is a minor groove-binding nonfluorescent quencher, and were used at concentrations of 300 nM and 500 nM, respectively. Each sample-reagent mix was partitioned into 765 reaction wells on a BiomarkTM 12.765 Digital Array using the NanoflexTM IFC Controller (Fluidigm). After partitioning, the array was placed in the BiomarkTM Real-time PCR System (Fluidigm) for thermal amplification and fluorescence detection. The reaction was initiated at 50oC for 2 min and continued at 95oC for 5 min followed by 45 cycles of 95oC for 15 sec and 59 oC for 1 min. After amplification, the number of informative wells (one that was positive for either the A or G allele only) and the number of wells positive for both alleles were counted and subjected to sequential probability ratio test (SPRT) analysis.

[0196] For a heterozygous euploid fetus, the A and G alleles should be equally represented (1:1) in the fetal genome, whereas for trisomy 18, there would be an additional copy of one allele thus giving a ratio of 2:1 in the fetal genome. A series of SPRT curves were generated for interpretation of different samples. These curves illustrate the expected proportion of informative wells positive for the overrepresented allele P_r (y-axis) for a given total number of informative wells (x-axis) needed for classification. For each sample, the experimentally derived P_r was compared with the expected P_r value. Samples above the upper curve were

classified as trisomy 18, whereas those below the bottom curve were classified as euploid. The area between the two curves is the unclassifiable region.

[0197] The feasibility of digital RNA-SNP analysis for the detection of fetal trisomy 18 was demonstrated by using the rs6098 SNP on the SERPINB2 gene. Placental tissue DNA samples from subjects with euploid and trisomy 18 fetuses were first genotyped by mass spectrometry for identifying heterozygous cases. Nine euploid and three trisomy 18 heterozygous placentas were found and subjected to digital RNA-SNP analysis. For each sample, P_r and m_r were calculated, and the SPRT curve for this m_r value was used for disease classification. As shown in FIG. 16A, all samples were correctly classified. The P_r values for trisomy 18 placentas were above the unclassifiable region, whereas those for euploid placentas fell below this region.

[0198] Samples with SPRT curves based on $m_r = 0.1$, 0.2, and 0.3 are illustrated in FIG. 16B. These data suggest that the digital RNA-SNP method is a valuable diagnostic tool for trisomy 18 pregnancies. The two curves represent the boundaries for the unclassifiable region. Samples with data points above the upper curve were classified as aneuploid, whereas those with data points below the bottom curve were classified as euploid.

C. Digital RCD analysis

[0199] This example demonstrates the performance of digital RCD analysis using microfluidics-based digital PCR. A variant of this approach is illustrated here, by way of example but not by way of limitation to, using a Fluidigm BioMark™ System. This system can perform over 9000 digital PCRs per run.

[0200] Placental tissues, maternal blood cell and amniotic fluid samples were obtained from pregnant women carrying euploid or trisomy 21 (T21) fetuses. Placental DNA from 10 T21 and 10 euploid cases was each mixed with an equal amount of euploid maternal blood cell DNA, thus producing twenty 50% DNA mixtures. To ensure accurate fetal proportion in the mixture samples, the extracted DNA was first quantified by optical density (OD) measurement at 260 nm. They were then digitally quantified by the BioMark™ System (Fluidigm) using the 12.765 Digital Arrays (Fluidigm). The assay for quantifying the samples was the same as described below except that only the probe for the reference chromosome was used.

[0201] The chromosome dosage in the 50% DNA mixtures and amniotic fluid samples was determined by digital PCR analysis of a nonpolymorphic chromosome 21 locus relative to one located on chromosome 1. A 101-bp amplicon of a pair of paralogous loci on chromosome 21 and 1 was first co-amplified by forward primer 5'-GTTGTTCTGCAAAAAACCTTCGA-3' and reverse primer 5'-CTTGGCCAGAAATACTTCATTACCATAT-3'. Two chromosome-specific TaqMan probes were designed to distinguish between the chromosome 21 and 1 paralogs, and their sequences were 5'-(FAM)TACCTCCATAATGAGTAAA(MGBNFQ)-3' and 5'-(VIC)CGTACCTCTGTAATGTGTAA(MGBNFQ)-3', respectively. The use of paralogous loci was used here by way of example only. In other words, non-paralogous loci could also be used for such analysis.

[0202] In order to demonstrate the use of the digital RCD approach to detect trisomy 18 (T18), another assay targeting paralogous sequences on chromosome 21 and 18 was designed. A 128-bp amplicon of the paralogous loci on chromosome 21 and 18 was first co-amplified by forward primer 5'-GTACAGAAACCACAAACTGATCGG-3' and reverse primer 5'-GTCCAGGCTGTGGGCCT-3'. Two chromosome-specific TaqMan probes were designed to distinguish between the chromosome 21 and 18 paralogs, and their sequences were 5'-(FAM)AAGAGGCGAGGCAA(MGBNFQ)-3' and 5'-(VIC)AAGAGGACAGGCAAC(MGBNFQ)-3', respectively. The use of paralogous loci was used here by way of example only. In other words, non-paralogous loci could also be used for such analysis.

[0203] All experiments were carried out on the BioMark™ System (Fluidigm) using the 12.765 Digital Arrays (Fluidigm). The reaction for one panel was set up using 2X TaqMan Universal PCR Master Mix (Applied Biosystems) in a reaction volume of 10 μL. Each reaction contained 1X TaqMan Universal PCR Master Mix, 900 nM of each primer, 125 nM of each probe and 3.5 μL of a 50% placental/maternal blood cell DNA sample. The sample/assay mixture was loaded into the Digital Array by the NanoFlex™ IFC controller (Fluidigm). The reaction was carried out on the BioMark™ System for detection. The reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 40 cycles of 95°C for 15s and 57°C for 1 min.

[0204] The euploid and T21 50% placental/maternal blood cell DNA samples were analyzed on the digital arrays with the chr21/chr1 assay. For each sample, the number of

informative wells, comprising the ones positive for only chr21 or chr1 markers (but not both), was counted. The proportion of the overrepresented marker among all the informative wells (P_r) was determined. SPRT curves appropriate for the exact average reference template concentration per well (m_r) for any one of the digital PCR panels were applied to determine if the experimentally-obtained P_r indicated a euploid or T21 sample. Data were aggregated from extra panels for samples which remained unclassified until a decision could be made. As shown in FIG. 17, all 50% placental/maternal blood cell DNA samples were correctly classified using this approach with data ranging from one to four panels needed. A SPRT curve was also plotted to show the decision boundaries for correct classification, as shown in FIG. 18.

[0205] We further applied the RCD analysis on amniotic fluid samples obtained from 23 women carrying euploid fetuses and 6 women carrying T21 fetuses. Each sample was analyzed in a single 765-well reaction panel with the chr21/chr1 assay. FIG. 19 shows the SPRT classification summary. As shown in FIG. 19, all the 29 samples were classified correctly. The digital RCD method is thus an alternative approach for the detection of trisomies using microsatellite (Levett LJ, et al. A large-scale evaluation of amnio-PCR for the rapid prenatal diagnosis of fetal trisomy. Ultrasound Obstet Gynecol 2001; 17: 115-8) or single nucleotide polymorphism (SNP) (Tsui NB, et al. Detection of trisomy 21 by quantitative mass spectrometric analysis of single-nucleotide polymorphisms. Clin Chem 2005; 51: 2358-62) markers or real-time non-digital PCR (Zimmermann B, et al. Novel real-time quantitative PCR test for trisomy 21. Clin Chem 2002; 48: 362-3) in miscellaneous sample types used for prenatal diagnosis, such as amniotic fluid and chorionic villus biopsies.

[0206] In an attempt to detect T18 cases, we applied the chr21/chr18 assay on 3 euploid and 5 T18 placental DNA samples. The proportion of the overrepresented marker among all the informative wells (P_r) was calculated. All of them were classified correctly except one T18 case was misclassified as euploid. The results were summarized in FIG. 20.

V. USING MULTIPLEX DIGITAL RCD ASSAYS ON MASS-SPECTROMETRIC PLATFORM

[0207] The number of plasma DNA molecules that are present per unit volume of maternal plasma is limited (Lo YMD. *et al.* 1998 *Am J Hum Genet* 62, 768-7758). For example, in early pregnancy, the median maternal plasma concentration of an autosomal locus, the *β-globin* gene, has been shown to be 986 copies/mL, with contributions from both the fetus

and mother (Lo YMD. *et al.* 1998 *Am J Hum Genet* 62, 768-7758). To capture 7,680 molecules, DNA extracted from some 8 mL of maternal plasma would be needed. This volume of plasma, obtainable from ~15 mL of maternal blood, is at the limit of routine practice. However, we envision that multiple sets of chr21 and reference chromosome targets can be combined for digital RCD analysis. For 5 pairs of chr21 and reference chromosome targets, just 1.6 mL of maternal plasma would be needed to provide the number of template molecules needed for analysis. Multiplex single molecule PCR could be performed. The robustness of such multiplex single molecule analysis has been demonstrated previously for single molecule haplotyping (Ding, C. and Cantor, CR. 2003 *Proc Natl Acad Sci U S A* 100, 7449-7453).

[0208] In one example, placental tissues and maternal blood cell samples were obtained from pregnant women carrying euploid or trisomy 21 (T21) fetuses. 5 euploid and 5 T21 placental DNA samples were each mixed with equal proportions of maternal blood cell DNA to produce 10 DNA mixtures mimicking plasma samples with 50% fetal DNA. To ensure accurate fetal proportion in the mixture samples, the extracted DNA was first quantified by optical density (OD) measurement at 260 nm. They were then digitally quantified by real-time PCR in 384-well format. The assay for quantifying the samples was the same as described in the previous example of digital RCD analysis.

[0209] The chromosome dosage in the 50% mix was determined by digital PCR analysis of a nonpolymorphic chromosome 21 locus relative to one located on chromosome 1. The method is called Digital Relative Chromosome Dosage (RCD) analysis. A 121-bp amplicon (inclusive of a 10-mer on each primer) of a pair of paralogous loci on chromosome 21 and 1 was co-amplified by forward primer 5'-ACGTTGGATGGTTGTTCTGCAAAAAACCTTCGA-3' and reverse primer 5'-ACGTTGGATGCTTGGCCAGAAATACTTCATTACCATAT-3'. An extension primer which targets the base differences between chromosome 21 and chromosome 1 was designed, and its sequence is 5'-CTCATCCTCACTTCGTACCTC-3'.

[0210] In order to demonstrate the utility of multiplexing digital PCR assays to detect T21 cases, another digital RCD assay targeting paralogous sequences on chromosome 21 and 18 was designed. A 148-bp amplicon (inclusive of a 10-mer on each primer) of the paralogous loci on chromosome 21 and 18 was co-amplified by forward primer 5'-ACGTTGGATGGTACAGAAACCACAAACTGATCGG-3' and reverse primer 5'-

ACGTTGGATGGTCCAGGCTGTGGGCCT-3'. An extension primer which targets the base differences between chromosome 21 and chromosome 18 was designed, and its sequence is 5'-ACAAAAGGGGGAAGAGG-3'.

[0211] Multiplex digital RCD analysis was performed using primer extension protocol. PCR reaction was set up using GeneAmp PCR Core Reagent Kit (Applied Biosystems) in a reaction volume of 5 μL. Each reaction contained 1X Buffer II, 2 mM $MgCl_2$, 200 μM dNTP mix, 0.2 U AmpliTaq Gold, 200 nM of each of the 4 primers and the 50% DNA mix. The assay/sample mixture was dispensed into 384-well PCR plate and the reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 40 cycles of 95°C for 15s and 57°C for 1 min.

[0212] PCR products were subjected to shrimp alkaline phosphatase (SAP) treatment to remove unincorporated dNTPs. The mixture was incubated at 37°C for 40 min followed by 85°C for 5 min. Primer extension reaction was then carried out. In brief, 771 nM of extension primer from chr21/chr1 assay, 1.54 μM of extension primer from chr21/chr18 assay, 0.67 U Thermosequenase (Sequenom), and 64 μM each of ddCTP, ddGTP, dATP and dTTP in an extension cocktail were added to the SAP-treated PCR products. The reaction conditions were 94°C for 2 min, followed by 94°C for 5 s, 50°C for 5 s, and 72°C for 5 s for 80 cycles. 16 μL of water and 3 mg of the Clean Resin (Sequenom) were added to the extension products for a final clean up. The mixtures were mixed in a rotator for 20 to 30 min, followed by a centrifugation step at 361 g for 5 min. Fifteen to 25 nL of the final products were dispensed onto a SpectroCHIP (Sequenom) by a MassARRAY Nanodispenser S (Sequenom). Data acquisition from the SpectroCHIP was done in the MassARRAY Analyzer Compact Mass Spectrometer (Sequenom). Mass data were imported into the MassARRAY Typer (Sequenom) software for analysis.

[0213] The five euploid and five T21 50% placental/maternal DNA samples were analyzed with the duplex RCD assay. For each sample, the number of informative wells from individual assay, comprising the ones positive for only chr21 or chr1 or chr18 markers, was counted. The proportion of the chr21 marker among all the informative wells (P_r) was calculated separately for each RCD assay. Sequential probability ratio test (SPRT) was then applied to determine if the P_r indicated a euploid or T21 sample. By doing so, the number of wells required was reduced as each plate was counted twice.

[0214] The chr21/chr1 assay was usually applied first. If the sample remained unclassified, then the values from the chr21/chr18 assay would be added for further calculations. Extra plates were used for samples which remained unclassified until a decision could be made. As shown in FIG. 21, all euploid 50% mix samples were correctly classified using a single 384-well plate. Several T21 cases required 2 or more plates for correct classification. If only one assay was used, a greater number of plates would be needed to attain the number of informative wells required when classification was achieved. For example, data for case N0230 was unclassifiable when either of the RCD assays was used alone. However, correct classification was achieved when data from the two assays were combined. If the duplex RCD assays were not used, additional plates of analyses would be needed. We would expect a further reduction of well number with a higher level of multiplexing of assays.

[0215] In another example, we developed a 4-plex assay targeting 4 different amplicons on chromosome 21 and their corresponding paralogous partners located on other non-chromosome-21 autosomes. This 4-plex assay was used in digital RCD analysis followed by SPRT classification of samples from euploid and trisomy 21 pregnancies. DNA extractions from placental samples were performed using the QIAamp tissue kit (Qiagen, Hilden, Germany).

[0216] All placental and maternal buffy coat DNA samples used in this study were first quantified by the NanoDrop spectrophotometer (NanoDrop Technology, Wilmington, DE). The DNA concentration was converted to genome equivalent (GE)/μL using a conversion of 6.6 pg/cell. The amount of DNA corresponding to approximately one template per well was determined by serially diluting the DNA samples. Under such a condition, we would expect approximately 37% of the wells to show negative amplification. In multiplex digital RCD analysis, 4 sets of paralogous sequence targets were selected: the paralogous loci on chromosome 21 and 1 were co-amplified by the forward primer 5'-ACGTTGGATGTTGATGAAGTCTCATCTCTACTTCG -3' and the reverse primer 5'-ACGTTGGATGCAATAAGCTTGGCCAGAAATACT -3', resulting in an amplicon of 81 bp. The paralogous loci on chromosome 21 and 7 were co-amplified by the forward primer 5'- ACGTTGGATGGAATTTAAGCTAAATCAGCCTGAACTG-3' and the reverse primer 5'- ACGTTGGATGGTTTCTCATAGTTCATCGTAGGCTTAT -3', resulting in an amplicon of 82 bp. The paralogous loci on chromosome 21 and 2 were co-amplified by the forward primer 5'- ACGTTGGATGTCAGGCAGGGTTCTATGCAG -3' and the reverse primer 5'-ACGTTGGATGAGGCGGCTTCCTGGCTCTTG -3', resulting in an amplicon of 101 bp.

The paralogous loci on chromosome 21 and 6 were co-amplified by the forward primer 5'-ACGTTGGATGGCTCGTCTCAGGCTCGTAGTT -3' and the reverse primer 5'-ACGTTGGATGTTTCTTCGAGCCCTTCTTGG -3', resulting in an amplicon of 102 bp. Each reaction contained 10X buffer II (Applied Biosystems), $MgCl_2$ and 100 nM of each primer. The total reaction volume was 5 μL/well. The reaction was initiated at 95°C for 5 min, followed by 45 cycles of 95°C for 30 sec, 62 °C for 30 sec and 72°C for 30 sec, and a final extension at 72 °C for 7 min. All conventional PCR amplifications were carried out on a GeneAmp PCR System 9700 (Applied Biosystems). The unincorporated nucleotides were deactivated by shrimp alkaline phosphatase (SAP) treatment. Each reaction contained 10X SAP buffer (Sequenom) and SAP enzyme (Sequenom). 2 μL of SAP mix was added to each PCR. The SAP reaction was incubated at 37 °C for 40 min and 85 °C for 5 min. After the SAP treatment, primer extension reaction was carried out on the PCR products using the iPLEX Gold kit (Sequenom). . The paralogous sequence mismatches (PSMs) on the paralogous loci on chromosomes 21 and 1 were interrogated by the extension primer 5'-GTCTCATCTCTACTTCGTACCTC -3'. The PSMs on the paralogous loci on chromosomes 21 and 7 were interrogated by the extension primer 5'-TTTTACGCTGTCCCCATTT-3'. The PSMs on the paralogous loci on chromosomes 21 and 2 were interrogated by the extension primer 5'-GGTCTATGCAGGAGCCGAC-3'. The PSMs on the paralogous loci on chromosomes 21 and 6 were interrogated by the extension primer 5'-TGGGCGCGGGAGCGGACTTCGCTGG-3'. Each reaction contained 10X iPLEX buffer (Sequenom), iPLEX termination mix (Sequenom), iPLEX enzyme (Sequenom) and 343 nM of each extension primer, except for the extension primer for PSMs on chromosomes 21 and 6 which was used at 1.03 μM. 2 μL of iPLEX mix was added to 5 μL of PCR product. The iPLEX reaction was cycled according to a 200-short-cycle program. Briefly, the samples were first denatured at 94 °C for 35 sec, followed by annealing at 52° C for 5 sec and extension at 80°C for 5 sec. The annealing and extension cycle was repeated four more times for a total of five cycles and then looped back to a 94° C denaturing step for 5 sec, after which was the 5-cycle annealing and extension loop again. The five annealing and extension cycles with the single denaturing step were repeated 39 times for a total of 40. A final extension at 72° C for 3 min was performed. The iPLEX reaction products were diluted with 16 μL water and desalted by 6 mg resin for each PCR. The 384-well plate was centrifuged at 1600 *g* for 3 min before dispensing onto the SpectroCHIP (Sequenom) and the matrix-assisted laser

desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry MS Analysis (Sequenom).

[0217] The number of wells which were positive for only chromosome 21 or only the reference chromosome for each of the four assays was independently recorded. For each assay, the Poisson corrected numbers of molecules for chromosome 21 and the reference chromosome were calculated. The sum of the Poisson corrected number of molecules for chromosome 21 as well as the sum of the Poisson corrected number of reference chromosomes from all four assays were calculated and deemed as the informative counts for the 4-plex assay. The P_r value was the chromosome 21 count for the 4-plex assay divided by the sum of the chromosome 21 and reference chromosome counts for the 4-plex assay. The experimentally derived P_r values were subjected to SPRT analysis. One or more 384-well plates would be analyzed until disease classification was possible by SPRT. A total of two 50% euploid placental genomic DNA/50% maternal buffy coat DNA mix and two 50% trisomy 21 placental genomic DNA/50% maternal buffy coat DNA mix were analyzed.

[0218] The experimentally derived P_r value would be compared with the expected value of P_r to test the null or alternative hypotheses. Alternatively, neither the null or alternative hypothesis could be accepted if the P_r for the given number of informative counts has not yet reached the required level of statistical confidence for disease classification. These cases were deemed unclassifiable until more data were available.

[0219] The results and the SPRT classification of each sample are tabulated in FIGS. 22A and 22B. The two euploid samples required 2 and 5 384-well multiplex digital RCD analyses before SPRT classification could be reached. Data from none of the individual member of the 4-plex assay allowed disease classificaiton by SPRT. Both trisomy 21 samples were each correctly classified with just one 384-well multiplex digital RCD analysis. Similarly, data from none of the individual member of the 4-plex assay allowed disease classificaiton by SPRT. However, the composite counts from the 4-plex assay allowed correct SPRT classification. These data illustrated that by using multiplex digital RCD, the effective number of informative counts were substantially increased for a given number of digital PCR analyses performed as compared to the use of a single-plex digital RCD assay.

VI. USING DIGITAL EPIGENETIC RELATIVE CHROMOSOME DOSAGE

[0220] Here we outline an approach called digital epigenetic relative chromosome dosage (digital ERCD) in which epigenetic markers exhibiting a fetal-specific DNA methylation pattern, or other epigenetic changes, on a chromosome involved in a chromosomal aneuploidy (e.g. chromosome 21 in trisomy 21) and on a reference chromosome, are subjected to digital PCR analysis. The ratio of the number of wells positive for the chromosome 21 epigenetic marker to that positive for the reference chromosome epigenetic marker in plasma DNA extracted from pregnant women bearing normal fetuses will give us the reference range. The ratio will be expected to be increased if the fetus has trisomy 21. It is obvious to those of skill in the art that more than one chromosome 21 markers and more than one reference chromosome markers could be used in this analysis.

[0221] One example of a gene on chromosome 21 which exhibits a fetus (placenta)-specific methylation pattern is the *Holocarboxylase synthetase (HLCS)* gene. *HLCS* is hypermethylated in the placenta, but hypomethylated in maternal blood cells; and is covered in U.S. patent application 11/784499, which is incorporated herein by reference. One example of a gene on a reference chromosome which exhibits a fetus (placenta)-specific methylation pattern is the *RASSF1A* gene on chromosome 3 [10]. *RASSF1A* is hypermethylated in the placenta but is hypomethylated in maternal blood cells, see U.S. patent application 11/784501, which is incorporated herein by reference.

[0222] In the application of hypermethylated *HLCS* and hypermethylated *RASSF1A* to digital PCR detection of trisomy 21 in a fetus using maternal plasma, maternal peripheral blood is first collected. Then the blood is subjected to centrifugation and the plasma is harvested. DNA from the plasma is then extracted using techniques well-known to those of skill in the art, such as using a QIAamp Blood kit (Qiagen). The plasma DNA is then subjected to digestion using one or more methylation-sensitive restriction enzymes, such as *Hpa*II and *Bst*UI. These methylation-sensitive restriction enzyme(s) will cut the maternal, nonmethylated versions of these genes, while leaving the fetal hypermethylated sequences intact. The digested plasma DNA sample is then diluted to an extent that on average approximately 0.2 to 1 molecule of either the restriction enzyme treated but intact *HLCS* or *RASSF1A* sequences will be detected per reaction well. Two real-time PCR systems will be used to amplify the diluted DNA, one with two primers and one TaqMan probe specific to the *HLCS* gene, encompassing the region that will be cut by the restriction enzyme(s) if the

sequence in unmethylated; and the other one towards the *RASSF1A* gene, similarly with two primers and one TaqMan probe. With regard to the latter *RASSF1A* primer/probe set, one example has been described by Chan et al 2006, Clin Chem 52, 2211-2218. The TaqMan probes towards the *HLCS* and *RASSF1A* targets will have different fluorescent reporters, such as FAM and VIC, respectively. A 384-well plate is then used to perform the digital PCR experiment. The number of wells scored positive for just *HLCS* and those scored positive for just *RASSF1A* will be counted, and a ratio of these counts will be taken. The *HLCS*:*RASSF1A* ratio will be expected to be higher for maternal plasma taken from a pregnant woman carrying a trisomy 21 fetus, when compared with one carrying a normal euploid fetus. The degree of overrepresentation will be dependent on the average reference template concentration per well in the digital PCR run.

[0223] Other methods for scoring these results will be possible, for the example the counting of the number of wells positive for *HLCS*, irrespective of the concurrent positivity for *RASSF1A*; and vice versa for *RASSF1A*, irrespective of the concurrent positivity for *HLCS*. Furthermore, in replacement of calculating the ratio, either the total number, or the difference in the *HLCS* and *RASSF1A* counts could be used to indicate the trisomy 21 status of a fetus.

[0224] Apart from doing the digital PCR in plates, it will also be obvious to those of skill in the art that other variants of digital PCR can be used, e.g. microfluidics chips, nanoliter PCR microplate systems, emulsion PCR, polony PCR and rolling-circle amplification, primer extension and mass spectrometry, etc. These variants of digital PCR are named by way of examples, and not as limitations.

[0225] Apart from real-time PCR, it will also be obvious to those of skill in the art that methods such as mass spectrometry can be used to score the digital PCR results.

[0226] Apart from using methylation-sensitive restriction enzymes to differentiate the fetal and maternal versions of *HLCS* and *RASSF1A*, it will be obvious to those of skill in the art that other methods for ascertaining the methylation status would also be applicable, e.g. bisulfite modification, methylation-specific PCR, immunoprecipitation using antibody to methylated cytosine, mass spectrometry, etc.

[0227] It will also be obvious to those of skill in the art that the approach illustrated in this example and other examples in this patent application can be used in the other bodily fluids in

which fetal DNA may be found, including maternal urine, amniotic fluid, transcervical washings, chorionic villus, maternal saliva, etc.

VII. MASSIVELY PARALLEL GENOMIC SEQUENCING USING EMULSION PCR AND OTHER STRATEGIES

[0228] Here we shall describe another example whereby a digital readout of nucleic acid molecules can be used for the detection of fetal chromosomal aneuploidies, e.g. trisomy 21, in maternal plasma. Fetal chromosomal aneuploidy results from abnormal dose(s) of a chromosome or chromosomal region. It is desirable that noninvasive tests have high sensitivity and specificity to minimize false diagnoses. However, fetal DNA is present in low absolute concentration and represent a minor portion of all DNA sequences in maternal plasma and serum. Hence, the number of digital PCR sampling targeting specific gene loci cannot be increased infinitely within the same biological specimen. Hence, the analysis of multiple sets of specific target loci may be used to increase the amount of data that could be obtained from a specimen without increasing the number of digital PCR sampling performed.

[0229] Accordingly, embodiments allow the noninvasive detection of fetal chromosomal aneuploidy by maximizing the amount of genetic information that could be inferred from the limited amount of fetal nucleic acids which exist as a minor population in a biological sample containing maternal background nucleic acids. In one aspect, the amount of genetic information obtained is sufficient for accurate diagnosis yet not overly excessive so as to contain costs and the amount of input biological sample required.

[0230] Massively parallel sequencing, such as that achievable on the 454 platform (Roche) (Margulies, M. *et al.* 2005 *Nature* 437, 376-380), Illumina Genome Analyzer (or Solexa platform) or SOLiD System (Applied Biosystems) or the Helicos True Single Molecule DNA sequencing technology (Harris TD et al. 2008 Science, 320, 106-109), the single molecule, real-time (SMRT™) technology of Pacific Biosciences, and nanopore sequencing (Soni GV and Meller A. 2007 Clin Chem 53: 1996-2001), allow the sequencing of many nucleic acid molecules isolated from a specimen at high orders of multiplexing in a parallel fashion (Dear Brief Funct Genomic Proteomic 2003; 1: 397-416). Each of these platforms sequences clonally expanded or even non-amplified single molecules of nucleic acid fragments.

[0231] As a high number of sequencing reads, in the order of hundred thousands to millions or even possibly hundreds of millions or billions, are generated from each sample in each run,

the resultant sequenced reads form a representative profile of the mix of nucleic acid species in the original specimen. For example, the haplotype, trascriptome and methylation profiles of the sequenced reads resemble those of the original specimen (Brenner et al Nat Biotech 2000; 18: 630-634; Taylor et al Cancer Res 2007; 67: 8511-8518). Due to the large sampling of sequences from each specimen, the number of identical sequences, such as that generated from the sequencing of a nucleic acid pool at several folds of coverage or high redundancy, is also a good quantitative representation of the count of a particular nucleic acid species or locus in the original sample.

[0232] In one embodiment, random sequencing is performed on DNA fragments that are present in the plasma of a pregnant woman, and one obtains genomic sequences which would originally have come from either the fetus or the mother. Random sequencing involves sampling (sequencing) a random portion of the nucleic acid molecules present in the biological sample. As the sequencing is random, a different subset (fraction) of the nucleic acid molecules (and thus the genome) may be sequenced in each analysis. Embodiments will work even when this subset varies from sample to sample and from analysis to analysis, which may occur even using the same sample. Examples of the fraction are about 0.1%, 0.5%, or 1% of the genome. In other embodiments, the fraction is at least any one of these values.

[0233] A bioinformatics procedure may then be used to locate each of these DNA sequences to the human genome. It is possible that a proportion of such sequences will be discarded from subsequent analysis because they are present in the repeat regions of the human genome, or in regions subjected to inter-individual variations, e.g. copy number variations. An amount of the chromosome of interest and of one or more other chromosomes may thus be determined.

[0234] In one embodiment, a parameter (e.g. a fractional representation) of a chromosome potentially involved in a chromosomal aneuploidy, e.g. chromosome 21 or chromosome 18 or chromosome 13, may then be calculated from the results of the bioinformatics procedure. The fractional representation may be obtained based on an amount of all of the sequences (e.g. some measure of all of the chromosomes) or a particular subset of chromosomes (e.g. just one other chromosome than the one being tested.)

[0235] This fractional representation is then compared to a reference range established in pregnancies involving normal (i.e. euploid) fetuses. It is possible that in some variants of the

procedure, the reference range would be adjusted in accordance with the fractional concentration of fetal DNA (f) in a particular maternal plasma sample. The value of f can be determined from the sequencing dataset, e.g. using sequences mappable to the Y chromosome if the fetus is male. The value of f may also be determined in a separate analysis, e.g. using fetal epigenetic markers (Chan KCA et al 2006 Clin Chem 52, 2211-8) or from the analysis of single nucleotide polymorphisms.

[0236] In one aspect, even when a pool of nucleic acids in a specimen is sequenced at <100% genomic coverage, and among the proportion of captured nucleic acid molecules, most of each nucleic acid species is only sequenced once, dosage imbalance of a particular gene locus or chromosome can also be quantitatively determined. In other words, the dosage imbalance of the gene locus or chromosome is inferred from the percentage representation of the said locus among all mappable sequenced tags of the specimen.

[0237] In one aspect for the massively parallel genomic sequencing approach, representative data from all of the chromosomes may be generated at the same time. The origin of a particular fragment is not selected ahead of time. The sequencing is done at random and then a database search may be performed to see where a particular fragment is coming from. This is contrasted from situations when a specific fragment from chromosome 21 and another one from chromosome 1 are amplified.

[0238] In one example, a proportion of such sequences would be from the chromosome involved in an aneuploidy such as chromosome 21 in this illustrative example. Yet other sequences from such a sequencing exercise would be derived from the other chromosomes. By taking into account of the relative size of chromosome 21 compared with the other chromosomes, one could obtain a normalized frequency, within a reference range, of chromosome 21-specific sequences from such a sequencing exercise. If the fetus has trisomy 21, then the normalized frequency of chromosome 21-derived sequences from such a sequencing exercise will increase, thus allowing the detection of trisomy 21. The degree of change in the normalized frequency will be dependent on the fractional concentration of fetal nucleic acids in the analyzed sample.

[0239] In one embodiment, we used the Illumina Genome Analyzer for single-end sequencing of human genomic DNA and human plasma DNA samples. The Illumina Genome Analyzer sequences clonally-expanded single DNA molecules captured on a solid surface termed a flow cell. Each flow cell has 8 lanes for the sequencing of 8 individual

specimens or pools of specimens. Each lane is capable of generating ~ 200Mb of sequence which is only a fraction of the 3 billion basepairs of sequences in the human genome. Each genomic DNA or plasma DNA sample was sequenced using one lane of a flow cell. The short sequence tags generated were aligned to the human reference genome and the chromosomal origin was noted. The total number of individual sequenced tags aligned to each chromosome were tabulated and compared with the relative size of each chromosome as expected from the reference human genome or non-disease representative specimens. Chromosome gains or losses were then identified.

[0240] The described approach is only one exemplification of the presently described gene/chromosome dosage strategy. Alternatively, paired end sequencing could be performed. Instead of comparing the length of the sequenced fragments from that expected in the reference genome as described by Campbell et al (Nat Genet 2008; 40: 722-729), the number of aligned sequenced tags were counted and sorted according to chromosomal location. Gains or losses of chromosomal regions or whole chromosomes were determined by comparing the tag counts with the expected chromosome size in the reference genome or that of a non-disease representative specimen

[0241] In another embodiment, the fraction of the nucleic acid pool that is sequenced in a run is further sub-selected prior to sequencing. For example, hybridization based techniques such as oligonucleotide array could be used to first sub-select for nucleic acid sequences from certain chromosomes, e.g. a potentially aneuploid chromosome and other chromosome(s) not involved in the aneuploidy tested. Another example is that a certain sub-population of nucleic acid sequences from the sample pool is sub-selected or enriched prior to sequencing. For example, it has been reported that fetal DNA molecules in maternal plasma are comprised of shorter fragments than the maternal background DNA molecules (Chan et al Clin Chem 2004; 50: 88-92). Thus, one may use one or more methods known to those of skill in the art to fractionate the nucleic acid sequences in the sample according to molecule size, e.g. by gel electrophoresis or size exclusion columns or by microfluidics-based approach. Yet, alternatively, in the example of analyzing cell-free fetal DNA in maternal plasma, the fetal nucleic acid portion could be enriched by a method that suppresses the maternal background, such as by the addition of formaldehyde (Dhallan et al JAMA 2004; 291: 1114-9).

[0242] Other single molecule sequencing strategies such as that by the Roche 454 platform, the Applied Biosystems SOLiD platform, the the Helicos True Single Molecule DNA

sequencing technology, the single molecule, real-time (SMRT™) technology of Pacific Biosciences, and nanopore sequencing could similarly be used in this application.

[0243] Examples of results and a further discussion (e.g. for sequencing and calculating parameters) may be found in concurrently filed application "DIAGNOSING FETAL CHROMOSOMAL ANEUPLOIDY USING GENOMIC SEQUENCING," (Attorney Docket No. 016285-005220US), which is incorporated by reference. Note that methods described herein for determining a cutoff value may be implemented when the reaction is sequencing, e.g. as described in this section.

[0244] The determination of the fractional concentration of fetal DNA in maternal plasma can also be done separate to the sequencing run. For example, the Y chromosome DNA concentration could be pre-determined using real-time PCR, microfluidics PCR or mass spectrometry. In fact, fetal DNA concentration could be determined using loci other than the Y chromosome and applicable to female fetuses. For example, Chan et al showed that fetal-derived methylated *RASSF1A* sequences would be detected in the plasma of pregnant women in the background of maternally derived unmethylated *RASSF1A* sequences (Chan et al, Clin Chem 2006;52:2211-8). The fractional fetal DNA concentration can thus be determined by dividing the amount of methylated *RASSF1A* sequences by the amount of total *RASSF1A* (methylated and unmethylated) sequences.

[0245] It is expected that maternal plasma would be preferred over maternal serum for practicing our invention because DNA is released from the maternal blood cells during blood clotting. Thus, if serum is used, it is expected that the fractional concentration of fetal DNA will be lower in maternal plasma than maternal serum. In other words, if maternal serum is used, it is expected that more sequences would need to be generated for fetal chromosomal aneuploidy to be diagnosed, when compared with a plasma sample obtained from the same pregnant woman at the same time.

[0246] Yet another alternative way of determining the fractional concentration of fetal DNA would be through the quantification of polymorphic differences between the pregnant women and the fetus (Dhallan R, et al. 2007 *Lancet*, 369, 474-481). An example of this method would be to target polymorphic sites at which the pregnant woman is homozygous and the fetus is heterozygous. The amount of fetal-specific allele can be compared with the amount of the common allele to determine the fractional concentration of fetal DNA.

[0247] In contrast to the existing techniques for detecting chromosomal aberrations, including comparative genomic hybridization, microarray comparative genomic hybridization, quantitative real-time polymerase chain reaction, which detect and quantify one or more specific sequence(s), massively parallel sequencing is not dependent on the detection or analysis of predetermined or a predefined set of DNA sequences. A random representative fraction of DNA molecules from the specimen pool is sequenced. The number of different sequence tags aligned to various chromosomal regions is compared between specimens containing or not containing tumoral DNA. Chromosomal aberrations would be revealed by differences in the number (or percentage) of sequences aligned to any given chromosomal region in the specimens.

[0248] In another example the sequencing technique on plasma cell-free DNA may be used to detect the chromosomal aberrations in the plasma DNA for the detection of a specific cancer. Different cancers have a set of typical chromosomal aberrations. Changes (amplifications and deletions) in multiple chromosomal regions may be used. Thus, there would be an increased proportion of sequences aligned to the amplified regions and a decreased proportion of sequences aligned to decreased regions. The percentage representation per chromosome could be compared with the size for each corresponding chromosome in a reference genome expressed as percentage of genomic representation of any given chromosome in relation to the whole genome. Direct comparisons or comparisons to a reference chromosome may also be used.

VIII. MUTATION DETECTION

[0249] Fetal DNA in maternal plasma exists as a minor population, with an average of 3% to 6% of maternal plasma DNA being contributed by the fetus. Because of this reason, most of the previous work in the field has focused on the detection of DNA targets which the fetus has inherited from the father, and which are distinguishable from the majority maternal DNA background in maternal plasma. Examples of such previously detected targets include the *SRY* gene on the Y chromosome (Lo YMD et al. 1998 *Am J Hum Genet,* 62, 768-775) and the *RHD* gene when the mother is RhD-negative (Lo YMD et al. 1998 *N Engl J Med,* 339, 1734-1738.

[0250] For fetal mutation detection, previous strategies using maternal plasma are limited to autosomal dominant conditions in which the father is a carrier, the exclusion of autosomal recessive diseases by direct mutation detection when the father and mother carries different

mutations, or by linkage analysis (Ding C. et al 2004 *Proc Natl Acad Sci U S A* 101, 10762-10767). These previous strategies have significant limitations. For example, for a couple where both the male and female partners carry the same mutation, then it would be impossible to carry out meaningful prenatal diagnosis by direct mutation detection in maternal plasma.

[0251] Such a scenario is illustrated in FIG. 23. In this scenario, there will be three possible fetal genotypes, NN, NM and MM, where N represents the normal allele and M represents the mutant allele. Examples of mutant alleles include those causing cystic fibrosis, beta-thalassemia, alpha-thalassemia, sickle cell anemia, spinal muscular atrophy, congenital adrenal hyperplasia, etc. Other examples of such disorders can be found in the Online Mendelian Inheritance in Man (OMIM) www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM&itool=toolbar. In maternal plasma, most of the DNA will be coming from the mother and would be NM. For any of the three fetal genotypes, there will not be any unique fetal allele which would allow its unique detection in maternal plasma. Thus, the conventional strategy cannot be applied here.

[0252] Embodiments described herein allow handing such scenarios. In the scenario where the mother and fetus are both NM, then the N allele and M allele will be in allelic balance. However, if the mother is NM and the fetus is NN, then there will be allelic imbalance in maternal plasma, with the N allele being overrepresented. On the other hand, if the mother is NM and the fetus is MM, then there will be allelic imbalance in maternal plasma, with the M allele being overrepresented. Thus, for fetal mutation detection, the null hypothesis refers to the absence of allelic imbalance when the fetus is of the NM genotype. The alternative hypothesis refers to the presence of allelic imbalance and the fetal genotype could be NN or MM depending on whether the N or M allele is overrepresented.

[0253] The presence or absence of allelic imbalance can be determined by digital PCR using embodiments described herein. In a first scenario, a particular volume of maternal plasma contains the DNA released from 100 cells, in which 50 are from the mother and 50 are from the fetus. Thus, the fractional concentration of fetal DNA in this volume of plasma is 50%. When the mother is of the genotype NM, then there will be 50 N alleles and 50 M alleles contributed by the mother. If the fetus is of the genotype NM, then there will be 50 N alleles and 50 M alleles contributed by the fetus. Therefore, there will be no allelic imbalance between the N allele and the M allele, a total of 100 copies each. On the other hand, if the

fetus is of the NN genotype, then there will be 100 fetal-derived N alleles in this volume of plasma. Thus, there will be a total of 150 N alleles to 50 M alleles. In other words, there will be allelic imbalance between N and M, with N being overrepresented at a ratio of 3:1 in relation to M.

[0254] In the converse situation, if the fetus is of the MM genotype, then there will be 100 fetal-derived M alleles in this volume of plasma. Thus, there will be 150 M alleles to 50 N alleles. In other words, there will be allelic imbalance between N and M, with M being overrepresented at a ratio of 3:1 in relation to N. Such allelic imbalance can be measured by digital PCR. The allele with the smaller number of positive wells is considered as the reference template. Similar to digital RNA-SNP and digital RCD analyses, the actual distribution of the alleles in the digital PCR experiment would be governed by the Poisson probability density function. Therefore, while the theoretical degree of allelic imbalance in the present scenario is 3:1, the expected degree of allelic imbalance would be dependent on the average template concentration per well during the digital PCR analysis. Thus interpretation cutoffs, such as for SPRT analysis, appropriate for the average reference template concentration per well (m_r) would need to be used for case classification.

[0255] Furthermore, the degree of allelic imbalance that needs to be measured is dependent on the fractional fetal DNA concentration. In contrast to the above example, let's consider a particular volume of maternal plasma contains the DNA released from 100 cells, in which 90 are from the mother and 10 are from the fetus. Thus, the fractional concentration of fetal DNA in this volume of plasma is 10%. When the mother is of the genotype NM, then there will be 90 N alleles and 90 M alleles contributed by the mother. If the fetus is of the genotype NM, then there will be 10 N alleles and 10 M alleles contributed by the fetus. Therefore, there will be no allelic imbalance between the N allele and the M allele, a total of 100 copies each. On the other hand, if the fetus is of the NN genotype, then there will be 20 fetal-derived N alleles in this volume of plasma. Thus, there will be a total of 110 N alleles to 90 M alleles.

[0256] In other words, there will be allelic imbalance between N and M, with N being overrepresented. In the converse situation, if the fetus is of the MM genotype, then there will be 20 fetal-derived M alleles in this volume of plasma. Thus, there will be 110 M alleles to 90 N alleles. In other words, there will be allelic imbalance between N and M, with M being overrepresented. The theoretical degree of allelic imbalance when the fetal DNA fractional concentration is 10% would be 110:90, which is different to the 3:1 ratio when there is 50%

fetal DNA as shown in the above example. Thus interpretation cutoffs, such as for SPRT analysis, appropriate for the fetal DNA fractional concentration would need to be used for case classification.

[0257] Thus, plasma DNA will be extracted. The amount of maternal and fetal DNA in the plasma sample will be quantified, for example by the real-time PCR assays previously established (Lo, *et al.* 1998 *Am J Hum Genet* 62, 768-775) or other types of quantifier well-known to those of skill in the art, e.g. SNP markers (Dhallan R et al. 2007 *Lancet,* 369, 474-481) and fetal epigenetic markers (Chan KCA et al. 2006 *Clin Chem,* 52, 2211-2218). The fetal DNA percentage will be calculated. Then the quantified plasma DNA sample is prepared (e.g. diluted or concentrated) such that during digital PCR analysis, each reaction well will contain an average of one template molecule (can be either the N or M allele). The digital PCR analysis will be carried out using a pair of primers, plus two TaqMan probes, one specific to the N allele, while the other one specific to the M allele. The number of wells which are positive only for M and the number of wells which are positive only for N will be counted. The ratio of these wells will be used to determine if there is evidence of allelic imbalance. Statistical evidence of allelic imbalance can be sought by methods well-known to those of skill in the art, such as using SPRT. In one variant of this analysis, it is also possible to count the number of wells which are positive either for M only or for M and N; as well as to count the number of wells which are positive either for N only or for M and N; and to derive a ratio of these counts. Once again, statistical evidence of allelic imbalance can be sought by methods well-known to those of skill in the art, such as using SPRT.

[0258] The dosage determination of fetal gene mutation, called digital relative mutant dosage (RMD), was validated using female/male (XX/XY) DNA mixtures. Blood cell DNA from a male and a female was each mixed with male DNA, producing samples with XX or XY genotypes in a background of XY at fractional concentrations of 25% and 50%, respectively, as shown in FIG. 24A.

[0259] In addition, blood cell samples were also obtained from 12 male and 12 female subjects. The female blood cell DNA (genotype XX) was each mixed with a 3-fold excess of male blood cell DNA (genotype XY), thus producing 12 DNA mixtures with 25% of DNA with XX genotype in a background of 75% DNA with XY genotype, with results shown in FIG. 24B.

[0260] An aim of the SPRT analysis was to determine the minor genotype present in the background DNA. In DNA mixtures with 25% of XX DNA in a background of 75% XY DNA, the minor allele would be the Y derived from the 75% of DNA. Since 25% of the DNA in the sample was of XX genotype, if there were a total of 200 molecules of DNA in the sample, then 150 molecules would have originated from the XY individual. Hence, the number of Y alleles would be expected to be 75. The number of X alleles contributed by the male proportion of DNA (genotype XY) is also 75. The number of X alleles contributed by the female (genotype XX) is 50 (2 times 25). Therefore, X to Y ratio is 125/75 = (1 +25%) / (1 – 25%) = 5/3.

[0261] For the second part of this study, blood cell samples were obtained from male and female subjects carrying HbE (G → A) and CD41/42 (CTTT/-) mutations on the *beta-globin* gene, i.e., the *hemoglobin, beta* (*HBB*) gene. To mimic maternal plasma samples obtained from heterozygous mothers (MN, where M= mutant and N= wildtype) bearing male fetuses with all possible genotypes (MM, MN or NN), blood cell DNA from males who were either homozygous for the wildtype alleles (NN) or heterozygous (MN) for one of the two mutations was each mixed with a blood cell DNA sample collected from females heterozygous for the same mutation (MN). DNA mixtures at various fractional male/mutant DNA concentrations were thus produced. Blood cell DNA sample from a female homozygous for the CD41/42 deletion (MM) was also used for preparing the DNA mixtures. To ensure an accurate male proportion for the SPRT classification, the fractional male DNA concentration of each DNA mixture was determined using the *ZFY/X* assay.

[0262] The digital *ZFY/X* assay was used for validating the SPRT as well as determining the fractional male DNA concentration in the DNA mixtures. The dosage of *Zinc Finger* protein sequences on chromosome X (*ZFX*) and Y (*ZFY*) was determined by digital PCR analysis. An 87-bp amplicon of the *ZFX* and *ZFY* loci was first co-amplified by the forward primer 5'-CAAGTGCTGGACTCAGATGTAACTG-3' and the reverse primer 5'-TGAAGTAATGTCAGAAGCTAAAACATCA-3'. Two chromosome-specific TaqMan probes were designed to distinguish between the chromosome X and Y paralogs, and their sequences were 5'-(VIC)TCTTTAGCACATTGCA(MGBNFQ)-3' and 5'-(FAM)TCTTTACCACACTGCAC(MGBNFQ)-3', respectively.

[0263] The mutant dosage in the DNA mixtures was determined by digital PCR analysis of the normal allele relative to the mutant allele. For the HbE mutation, an 87-bp amplicon of

the normal and mutant alleles was first co-amplified by the forward primer 5'-GGGCAAGGTGAACGTGGAT -3' and the reverse primer 5'-CTATTGGTCTCCTTAAACCTGTCTTGTAA -3'. Two allele-specific TaqMan probes were designed to distinguish between the normal (G) and mutant (A) alleles, and their sequences were 5'-(VIC) TTGGTGGTGAGGCC (MGBNFQ)-3' and 5'-(FAM) TTGGTGGTAAGGCC (MGBNFQ)-3', respectively. Results for the HbE mutation are shown in FIG. 25.

[0264] For the CD41/42 deletion mutation, an 87- and 83-bp amplicon of the normal and mutant alleles was first co-amplified by the forward primer 5'-TTTTCCCACCCTTAGGCTGC -3' and the reverse primer 5'-ACAGCATCAGGAGTGGACAGATC -3', respectively. Two allele-specific TaqMan probes were designed to distinguish between the normal (without deletion) and mutant (with deletion) alleles, and their sequences were 5'-(VIC) CAGAGGTTCTTTGAGTCCT (MGBNFQ)-3' and 5'-(FAM) AGAGGTTGAGTCCTT (MGBNFQ)-3', respectively. Results for the HbE mutation are shown in FIGS. 26A and 26B.

[0265] These experiments were carried out on the BioMark™ System (Fluidigm) using the 12.765 Digital Arrays (Fluidigm). The reaction for one panel was set up using 2X TaqMan Universal PCR Master Mix (Applied Biosystems) in a reaction volume of 10 μL. For the CD41/42 and *ZFY/X* assays, each reaction contained 1X TaqMan Universal PCR Master Mix, 900 nM of each primer, 125 nM of each probe and 3.5 μL of DNA mixture at 1 ng/μL. For the HbE assay, 250 nM and 125 nM of probes targeting the normal (G) and mutant (A) alleles were added, respectively. The sample/ assay mixture was loaded into the Digital Array by the NanoFlex™ IFC controller (Fluidigm). The reaction was carried out on the BioMark™ System for signal detection. The reaction was initiated at 50°C for 2 min, followed by 95°C for 10 min and 50 cycles of 95°C for 15s and 57°C (for *ZFY/X* and CD41/42) or 56°C (for HbE) for 1 min. At least one reaction panel was used for each case, and data were aggregated from extra panels for samples which remained unclassified until a decision could be made.

[0266] It will also be obvious to those of skill in the art that the digital PCR can be performed using methods well-known to those of skill in the art, e.g. microfluidics chips, nanoliter PCR microplate systems, emulsion PCR, polony PCR, rolling-circle amplification, primer extension and mass spectrometry.

IX. EXAMPLE WITH CANCER

[0267] In one embodiment, the present invention may be performed to classify a sample as having allelic ratio skewing or not, as may occur in a cancerous tumor. In one aspect, for each case, the number of wells with positive signal for the A allele only, the G allele only, and both alleles were determined by digital PCR. The reference allele was defined as the allele with the smaller number of positive wells. (In the unlikely scenario that both alleles have the same number of positive wells, then either can be used as the reference allele.) The inferred average concentration of the reference allele per well (m_r) was calculated using the total number of wells negative for the reference allele, irrespective whether the other allele was positive, according to the Poisson probability density function. We use a hypothetical example to illustrate the calculation.

[0268] In a 96-well reaction, 20 wells are positive for the A allele, 24 wells are positive for the G allele, and 28 wells are positive for both alleles. The A allele would be regarded as the reference allele because less wells are positive for this allele. The number of wells negative for the reference allele would be 96 - 20 - 28 = 48. Therefore, m_r can be calculated using the Poisson distribution and would be -1n(48/96) = 0.693.

[0269] In the context of LOH detection, the null hypothesis refers to a sample that is assumed to lack allelic ratio skewing caused by the presence of a deletion of one allele. Under this assumption, the expected ratio of the number of positive wells for the two alleles would be 1:1 and, thus, the expected proportion of informative wells (wells positive for only one allele) containing the potentially overrepresented allele would be 0.5.

[0270] In the context of LOH detection, the alternative hypothesis refers to a sample that is assumed to have allelic ratio skewing caused by the presence of a deletion of one allele in 50% of the cells of the sample. As the allelic ratio between the overrepresented allele and the reference allele is 2:1, the average concentration of the overrepresented allele per well would be doubled that of the reference allele. However, the number of wells positive for the overrepresented allele would not be simply two times that for the reference allele but would follow the Poisson distribution.

[0271] An informative well is defined as a well positive for either the A or the G allele but not for both alleles. The calculation of the expected proportion of the number of wells containing the overrepresented alleles for samples with allelic ratio skewing is the same as is shown in Table 600. In the above example, if LOH is present in 50% of tumor cells, the

average concentration of the G allele per well would be 2 times 0.693 = 1.386. If LOH is present in more than 50% of the tumor cells, then the average concentration of the G allele per well would be according to the formula: $1/[1-(\text{proportion with LOH})] \times m_r$.

[0272] The expected proportion of wells positive for the G allele would be $1 - e^{-1..386} = 0.75$ (i.e., 75% or 72 wells). Assuming that the positivity of a well for the A or G allele is independent, 0.5 x 0.75 = 0.375 of the wells would be positive for both the A and G alleles. Hence, 0.5 – 0.375 = 0.125 of the wells would be positive for the A allele only and 0.75 – 0.375 = 0.375 of the wells would be positive for the G allele only. Therefore, the proportion of informative wells would be 0.125 + 0.375 = 0.5. The expected proportion of informative wells carrying the G allele would be 0.375/0.5 = 0.75. This expected value for P_r can then be used for the construction of appropriate SPRT curves for determining whether allelic ratio skewing (i.e. LOH in this context) is present in the sample.

[0273] The actual proportion of informative wells carrying the non-reference allele experimentallly determined by the digital PCR analysis (P_r) was then used to determine whether the null or alternative hypothesis would be accepted, or whether further analysis with more wells would be necessary. The decision boundaries for P_r to accept the null or alternative hypothesis was calculated based on a threshold likelihood ratio of 8 as this value had been shown to provide satisfactory performance to discriminate samples with and without allelic imbalance in the context of cancer detection (Zhou, W, et al. (2001) *Nat Biotechnol* 19, 78-81; Zhou et a1 2002, supra). In the above example, the number of informative wells would be 20 +24 = 44 and the experimentally obtained P_r would be 24/44 = 0.5455. The decision boundaries would be ≤ 0.5879 to accept the null hypothesis and ≥ 0.6739 to accept the alternative hypothesis. Therefore, the sample in this example would be classified as NOT having allelic ratio skewing.

[0274] In conclusion, we outlined an approach to detect sequence imbalance in a sample. In one embodiment, this invention can be used for the noninvasive detection of fetal chromosomal aneuploidy, such as trisomy 21 by analysis of fetal nucleic acids in maternal plasma. This approach can also be applied to other biological materials containing fetal nucleic acids, including amniotic fluid, chorionic villus samples, maternal urine, endocervical samples, maternal saliva, etc. First, we demonstrated the use of this invention for determining allelic imbalance of a SNP on *PLAC4* mRNA, a placenta-expressed transcript on chromosome 21, in maternal plasma of women bearing trisomy 21 fetuses. Second, we

demonstrated that our invention can be used as a non-polymorphism based method, through relative chromosome dosage (RCD) analysis, for the noninvasive prenatal detection of trisomy 21. Such a digital RCD-based approach involves the direct assessment of whether the total copy number of chromosome 21 in a sample containing fetal DNA is overrepresented with respect to a reference chromosome. Even without elaborate instrumentation, digital RCD allows the detection of trisomy 21 in samples containing 25% fetal DNA. We applied the sequential probability ratio test (SPRT) to interpret the digital PCR data. Computer simulation analyses confirmed the high accuracy of the disease classification algorithm.

[0275] We further outlined that the approach can be applied to the determination of other forms of nucleic acid sequence imbalances other than chromosomal aneuploidy, such as that for the detection of fetal mutation or polymorphism detection in maternal plasma and regional gains and losses in the genomes of malignant cells through the analysis of tumor-derived nucleic acids in plasma.

[0276] Any of the software components or functions described in this application, may be implemented as software code to be executed by a processor using any suitable computer language such as, for example, Java, C++ or Perl using, for example, conventional or object-oriented techniques. The software code may be stored as a series of instructions, or commands on a computer readable medium for storage and/or transmission, suitable media include random access memory (RAM), a read only memory (ROM), a magnetic medium such as a hard-drive or a floppy disk, or an optical medium such as a compact disk (CD) or DVD (digital versatile disk), flash memory, and the like. The computer readable medium may be any combination of such storage or transmission devices.

[0277] Such programs may also be encoded and transmitted using carrier signals adapted for transmission via wired, optical, and/or wireless networks conforming to a variety of protocols, including the Internet. As such, a computer readable medium according to an embodiment of the present invention may be created using a data signal encoded with such programs. Computer readable media encoded with the program code may be packaged with a compatible device or provided separately from other devices (e.g., via Internet download). Any such computer readable medium may reside on or within a single computer program product (e.g. a hard drive or an entire computer system), and may be present on or within different computer program products within a system or network. A computer system may

include a monitor, printer, or other suitable display for providing any of the results mentioned herein to a user.

[0278] An example of a computer system is shown in FIG. 27. The subsystems shown in FIG. 27 are interconnected via a system bus 2775. Additional subsystems such as a printer 2774, keyboard 2778, fixed disk 2779, monitor 2776, which is coupled to display adapter 2782, and others are shown. Peripherals and input/output (I/O) devices, which couple to I/O controller 2771, can be connected to the computer system by any number of means known in the art, such as serial port 2777. For example, serial port 2777 or external interface 2781 can be used to connect the computer apparatus to a wide area network such as the Internet, a mouse input device, or a scanner. The interconnection via system bus allows the central processor 2773 to communicate with each subsystem and to control the execution of instructions from system memory 2772 or the fixed disk 2779, as well as the exchange of information between subsystems. The system memory 2772 and/or the fixed disk 2779 may embody a computer readable medium.

[0279] The above description of exemplary embodiments of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form described, and many modifications and variations are possible in light of the teaching above. The embodiments were chosen and described in order to best explain the principles of the invention and its practical applications to thereby enable others skilled in the art to best utilize the invention in various embodiments and with various modifications as are suited to the particular use contemplated.

[0280] All publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety for all purposes.

WHAT IS CLAIMED IS:

1. A method for determining whether a nucleic acid sequence imbalance exists within a biological sample, the method comprising:

receiving data from a plurality of reactions, wherein the data includes:

(1) a first set of quantitative data indicating a first amount of a clinically relevant nucleic acid sequence; and

(2) a second set of quantitative data indicating a second amount of a background nucleic acid sequence different from the clinically relevant nucleic acid sequence;

determining a parameter from the two data sets;

deriving a first cutoff value from an average concentration of a reference nucleic acid sequence in each of the plurality of reactions, wherein the reference nucleic acid sequence is either the clinically relevant nucleic acid sequence or the background nucleic acid sequence;

comparing the parameter to the first cutoff value; and

based on the comparison, determining a classification of whether a nucleic acid sequence imbalance exists.

2. The method of claim 1 wherein the first set of data are obtained from one or more first markers that each detect a presence of a part of the clinically relevant nucleic acid sequence in a reaction, and wherein the second set of data are obtained from one or more second markers that each detect a presence of a part of the background nucleic acid sequence in a reaction.

3. The method of claim 1, further comprising:

determining the average concentration of the reference nucleic acid sequence in each of the plurality of reaction using an inverse of a probability distribution having an input of a value derived from the data for the reference nucleic acid sequence.

4. The method of claim 1 wherein the clinically relevant nucleic acid sequence is from chromosome 21 and the background nucleic acid sequence is from a chromosome other than chromosome 21.

5. The method of claim 1 wherein the clinically relevant nucleic acid sequence is from chromosome 18 or 13 and the background nucleic acid sequence is from a chromosome other than chromosome 18 or 13 respectively.

6. The method of claim 1 wherein the clinically relevant nucleic acid sequence is an allele of a genetic polymorphism, and the background nucleic acid sequence is another allele of the genetic polymorphism.

7. The method of claim 1 wherein the clinically relevant nucleic acid sequence is a mutated copy of the cystic fibrosis transmembrane conductance regulator (CFTR) gene, the beta-globin gene or the alpha-globin gene, and the background nucleic acid sequence is from the wildtype copy of the corresponding gene.

8. The method of claim 1 wherein the biological sample is plasma or serum from a pregnant woman.

9. The method of claim 1 wherein a reaction is an amplification reaction.

10. The method of claim 9 wherein a reaction is a part of a digital PCR process.

11. The method of claim 1 wherein a reaction is a sequencing reaction.

12. The method of claim 1 wherein first portions of the clinically relevant nucleic acid sequence and the background nucleic acid sequence are from a first individual and second portions of the clinically relevant nucleic acid sequence and the background nucleic acid sequence are from a second individual.

13. The method of claim 12 wherein the cutoff value is based on a measurement of one of the first portions or a measurement of one of the second portions.

14. The method of claim 1, further comprising comparing the parameter to a second cutoff value.

15. The method of claim 14 wherein the classifications include disease state, non-disease state, and non-classifiable.

16. The method of claim 14 wherein the classifications include homozygous, heterozygous, and non-classifiable.

17. The method of claim 14 wherein the second cutoff value is based on a ratio of the first amount of the clinically relevant nucleic acid sequence relative to the second amount of the background nucleic acid sequence in a non-diseased state.

18. The method of claim 1 wherein the parameter is a ratio of the first amount of the clinically relevant nucleic acid sequence relative to the second amount of the background nucleic acid sequence.

19. The method of claim 1 wherein calculating the first cutoff value includes using at least one of: sequential probability ratio testing, false discovery rates, confidence intervals, and receiver operating characteristic curves.

20. The method of claim 1 wherein deriving the first cutoff value includes:

determining a proportion P1 of informative reactions containing an overrepresented nucleic acid sequence, which is either the reference or non-reference nucleic acid sequence; and

calculating the first cutoff value from the first proportion P1.

21. The method of claim 20 wherein determining the proportion P1 comprises:

determining a first probability of a reaction containing at least one of the overrepresented nucleic acid sequence;

calculating a second probability that a reaction is informative; and

calculating the proportion P1 using the first probability and the second probability.

22. The method of claim 21 wherein the first probability is determined by multiplying the average concentration of the reference nucleic acid sequence by an expected ratio relative to the non-reference nucleic acid sequence.

23. The method of claim 21 wherein the first probability is determined using the Poisson distribution having the average concentration of the overrepresented nucleic acid sequence in each of the plurality of reactions as an input.

24. The method of claim 21, further comprising:

determining a third probability of a reaction containing at least one of the underrepresented nucleic acid sequence, wherein calculating the second probability that the reaction is informative includes assuming that first probability and the second probability are independent.

25. A method for determining whether a nucleic acid sequence imbalance exists within a biological sample, the method comprising:

receiving data from a plurality of reactions, wherein the data includes:

(1) a first set of quantitative data indicating a first amount of a clinically relevant nucleic acid sequence; and

(2) a second set of quantitative data indicating a second amount of a background nucleic acid sequence different from the clinically relevant nucleic acid sequence, wherein the clinically relevant nucleic acid sequence and the background nucleic acid sequence come from a first type of cells and from one or more second types of cells;

determining a parameter from the two data sets;

deriving a first cutoff value from a first percentage resulting from a measurement of an amount of a nucleic acid sequence from the first type of cells in the biological sample;

comparing the parameter to the first cutoff value; and

based on the comparison, determining a classification of whether a nucleic acid sequence imbalance exists.

26. The method of claim 25 wherein the first type of cells is from a first organism and the second types of cells are from a second organism.

27. The method of claim 25 wherein deriving the first cutoff value includes:

determining a first average concentration of a reference nucleic acid sequence per reaction, wherein the reference nucleic acid sequence is the clinically relevant nucleic acid sequence or the background nucleic acid sequence that is underrepresented; and

multiplying the first average concentration by a factor derived from the first percentage to obtain a second average concentration of the nucleic acid sequence that is not the reference nucleic acid sequence.

28. The method of claim 27, further comprising:

determining the average concentration of the reference nucleic acid sequence in each of the plurality of reactions using an inverse of a probability distribution having an input of a value derived from the data for the reference nucleic acid sequence.

29. The method of claim 28 wherein the probability distribution is the Poisson distribution.

30. The method of claim 25 wherein the percentage is measured by determining an amount of a fetal-specific marker using quantitative real-time PCR, digital PCR, semiquantitative competitive PCR, real-competitive PCR, or mass spectrometry.

31. A computer program product comprising a computer readable medium encoded with a plurality of instructions for controlling a computing system to perform an operation for determining whether a nucleic acid sequence imbalance exists within a biological sample, the operation comprising the steps of:

receiving data from a plurality of reactions, wherein the data includes:

(1) a first set of quantitative data indicating a first amount of a clinically relevant nucleic acid sequence; and

(2) a second set of quantitative data indicating a second amount of a background nucleic acid sequence different from the clinically relevant nucleic acid sequence;

determining a parameter from the two data sets;

deriving a first cutoff value from an average concentration of a reference nucleic acid sequence in each of the plurality of reactions, wherein the reference nucleic acid sequence is either the clinically relevant nucleic acid sequence or the background nucleic acid sequence;

comparing the parameter to the first cutoff value; and

based on the comparison, determining a classification of whether a nucleic acid sequence imbalance exists.

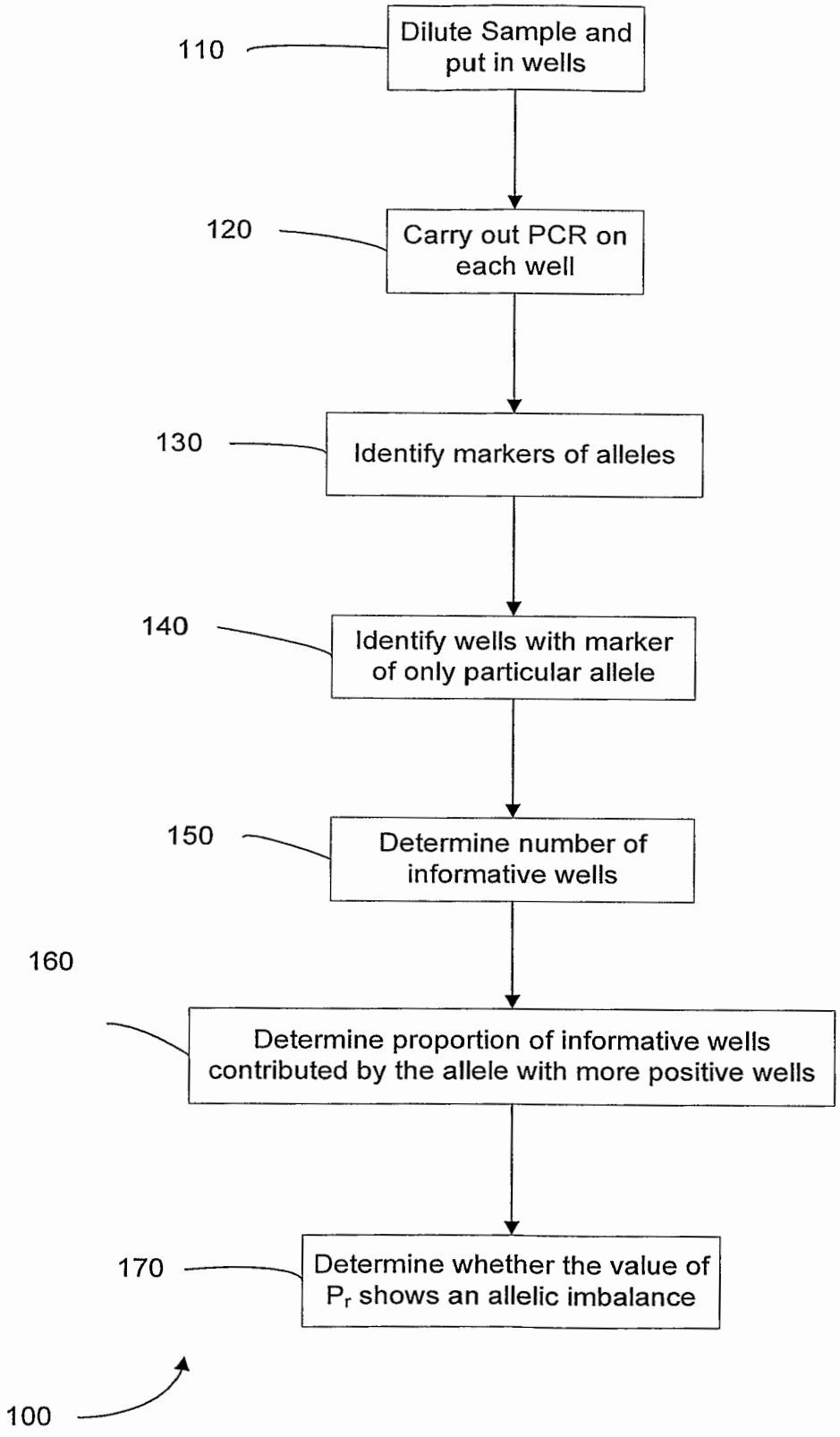

110
Dilute Sample and put in wells
120
Carry out PCR on each well
130
Identify markers of alleles
140
Identify wells with marker of only particular allele
150
Determine number of informative wells
160
Determine proportion of informative wells contributed by the allele with more positive wells
170
Determine whether the value of P_r shows an allelic imbalance
100

FIG. 1

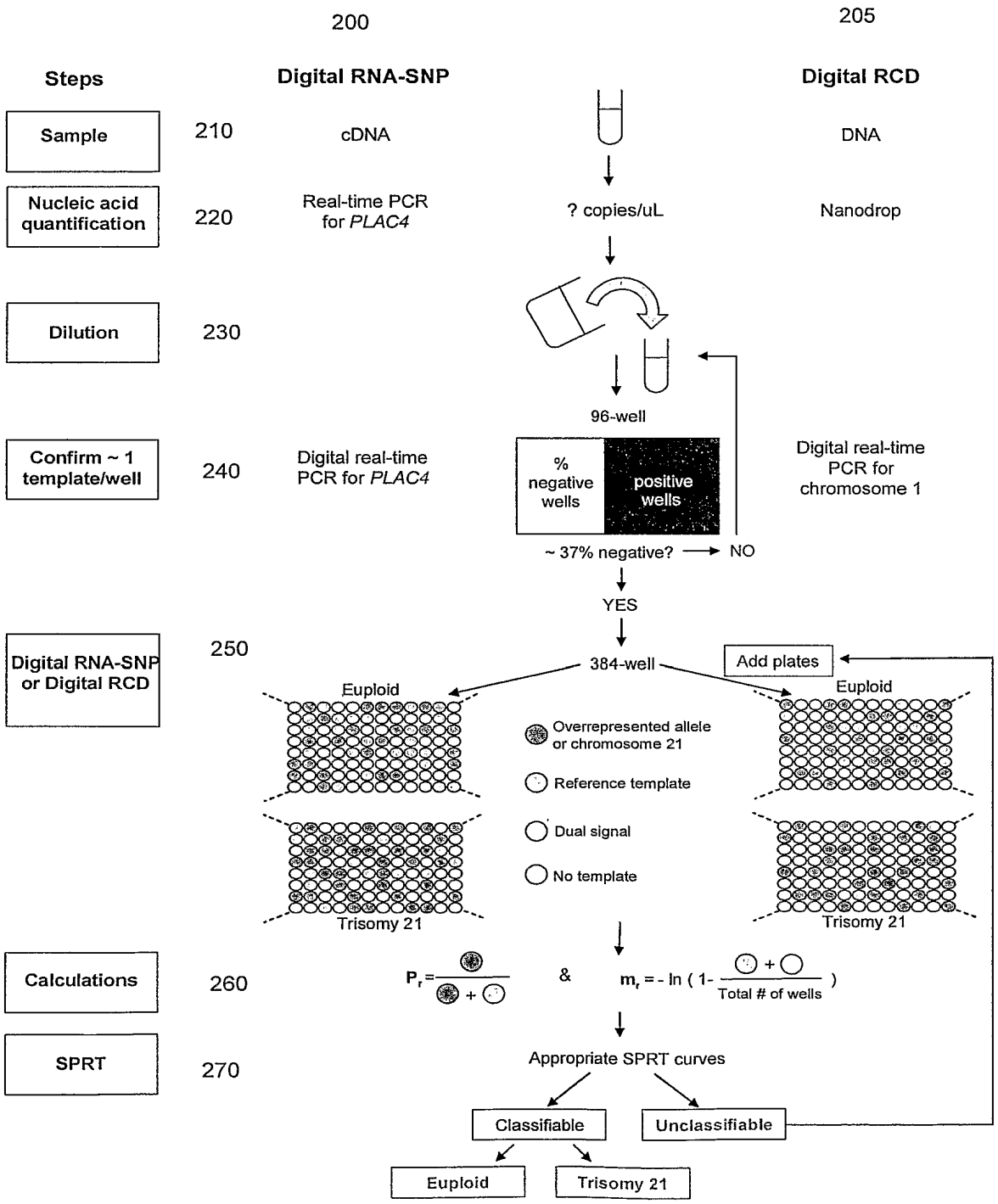

200
205
Steps
Digital RNA-SNP
Digital RCD
Sample
210
cDNA
DNA
Nucleic acid quantification
220
Real-time PCR for PLAC4
? copies/uL
Nanodrop
Dilution
230
96-well
Confirm ~ 1 template/well
240
Digital real-time PCR for PLAC4
% negative wells
positive wells
Digital real-time PCR for chromosome 1
~ 37% negative?
NO
YES
Digital RNA-SNP or Digital RCD
250
384-well
Add plates
Euploid
Euploid
Overrepresented allele or chromosome 21
Reference template
Dual signal
No template
Trisomy 21
Trisomy 21
Calculations
260
$P_r = \frac{\text{Overrepresented}}{\text{Overrepresented} + \text{Reference}}$
&
$m_r = -\ln\left(1 - \frac{\text{Reference} + \text{Dual}}{\text{Total \# of wells}}\right)$
SPRT
270
Appropriate SPRT curves
Classifiable
Unclassifiable
Euploid
Trisomy 21

FIG. 2A

Cancer	Common chromosomal aberrations	
	Gain	Loss
Bladder cancer	7	9p; 11q; 17p
Breast cancer	1q; 8q; 17q; 20q	8p; 13q; 16q; 18q
Colorectal cancer	7; 8q; 13q; 20q	1p; 5q; 8p; 15q; 17p; 18q
Hepatocellular carcinoma	1q; 8q; 20q	4q; 8p; 13q; 16q; 17p
Non - small cell lung carcinoma	1q; 3q; 5p; 8q	3p; 8p; 9p; 13q; 17p
Small cell lung cancinoma	3q; 5p; 8q; 19q	3p; 4q; 5q; 10q; 13q; 16q
Nasopharyngeal carcinoma	1q; 3q; 8; 12; 19	1p; 3p; 9; 11q; 13q; 14q; 16q
Prostate cancer	7; 8q; Xq	8p; 13q; 16q

FIG. 2B

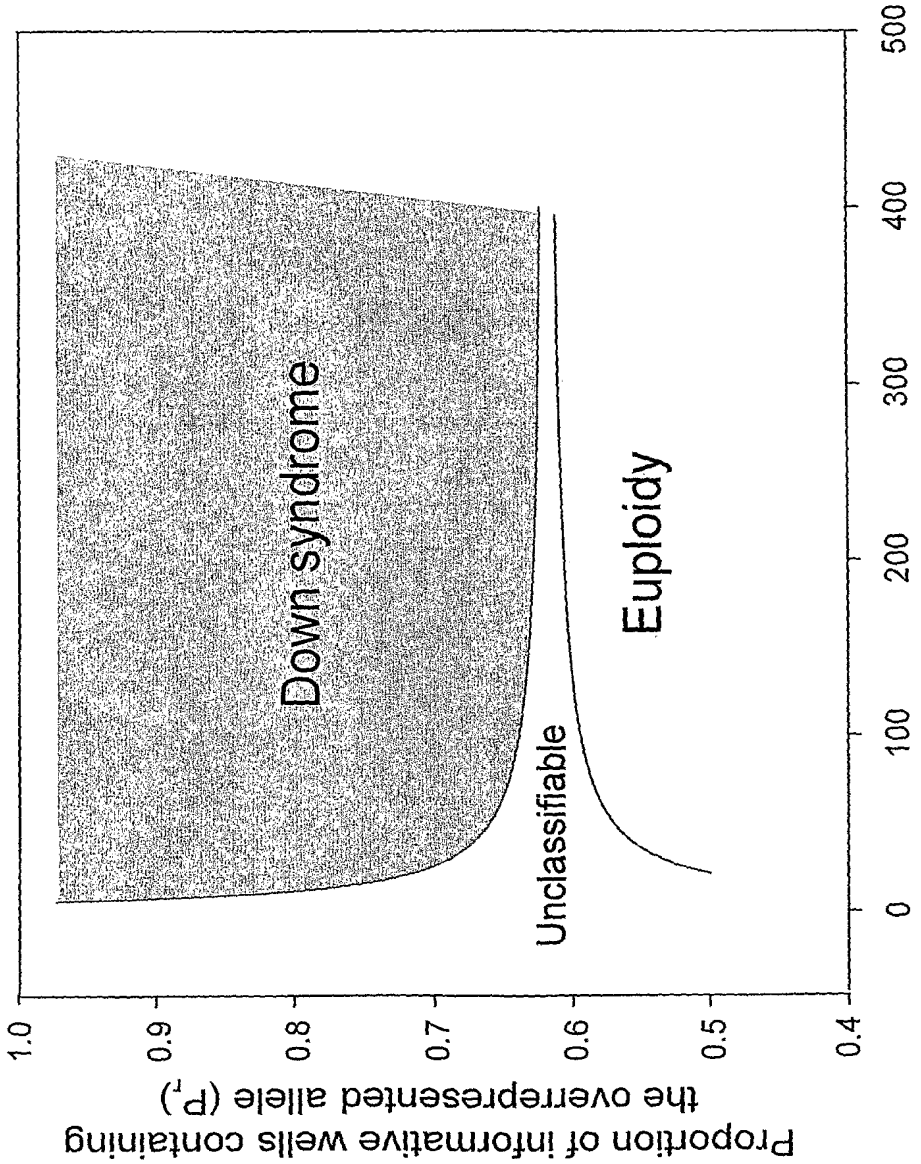

Down syndrome
Unclassifiable
Euploidy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
100
200
300
400
500
No. of informative wells
Proportion of informative wells containing the overrepresented allele (P_r)

FIG. 3

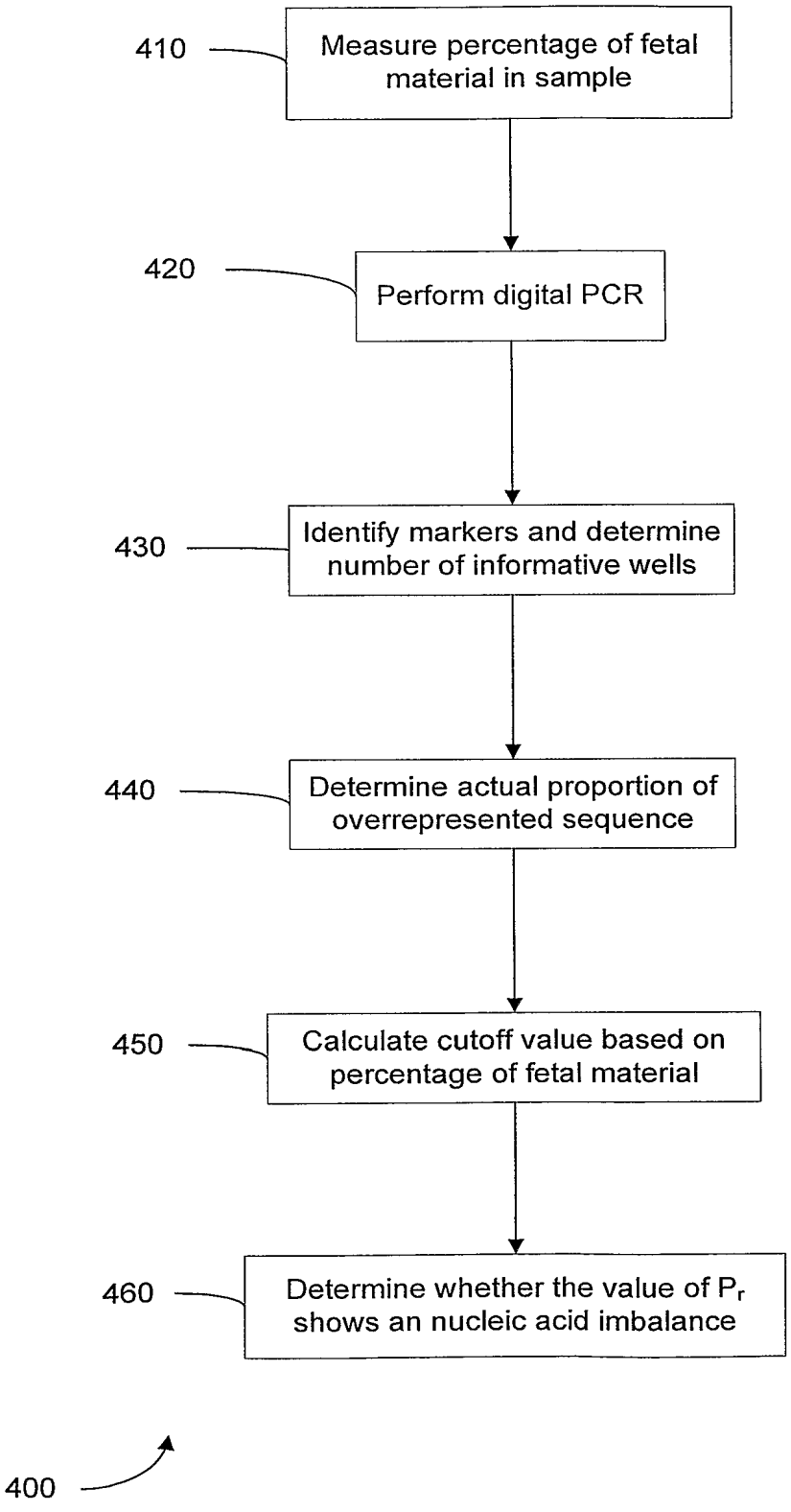

410
Measure percentage of fetal material in sample
420
Perform digital PCR
430
Identify markers and determine number of informative wells
440
Determine actual proportion of overrepresented sequence
450
Calculate cutoff value based on percentage of fetal material
460
Determine whether the value of P_r shows an nucleic acid imbalance
400

FIG. 4

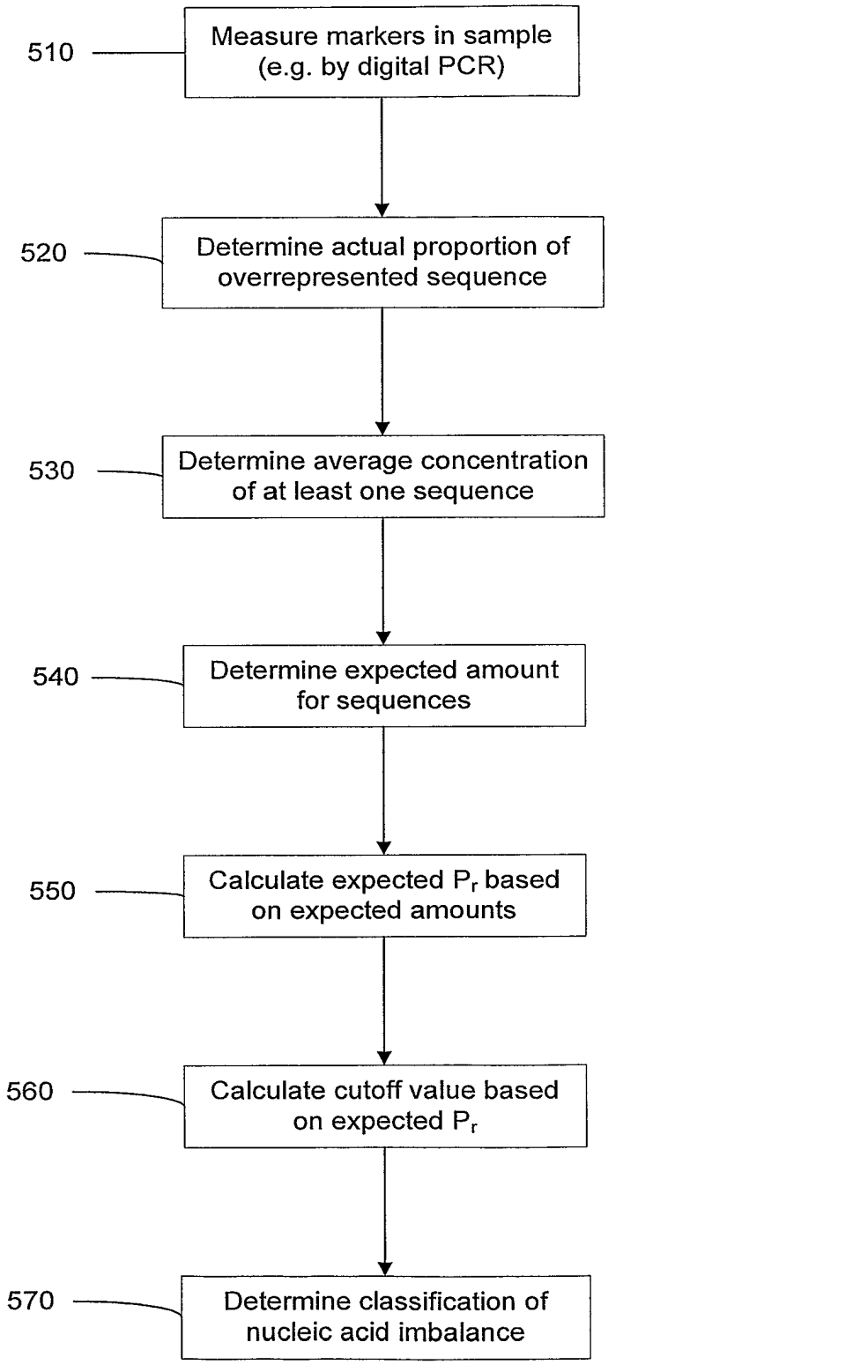

510
Measure markers in sample (e.g. by digital PCR)
520
Determine actual proportion of overrepresented sequence
530
Determine average concentration of at least one sequence
540
Determine expected amount for sequences
550
Calculate expected P_r based on expected amounts
560
Calculate cutoff value based on expected P_r
570
Determine classification of nucleic acid imbalance
500

FIG. 5

m_r	Proportion of wells positive for the reference allele	Proportion of wells positive for the overrepresented allele	Proportion of wells positive for both alleles	Proportion of wells positive for the reference allele ONLY	Proportion of wells positive for the overrepresented allele ONLY	Digital RNA-SNP allelic ratio	Proportion of informative wells	P_r
0.1	0.0952	0.1813	0.0173	0.0779	0.164	2.11	0.2419	0.68
0.2	0.1813	0.3297	0.0598	0.1215	0.2699	2.22	0.3914	0.69
0.3	0.2592	0.4512	0.1169	0.1422	0.3342	2.35	0.4765	0.7
0.4	0.3297	0.5507	0.1815	0.1481	0.3691	2.49	0.5173	0.71
0.5	0.3935	0.6321	0.2487	0.1448	0.3834	2.65	0.5282	0.73
0.6	0.4512	0.6988	0.3153	0.1359	0.3835	2.82	0.5194	0.74
0.7	0.5034	0.7534	0.3793	0.1241	0.3741	3.01	0.4983	0.75
0.8	0.5507	0.7981	0.4395	0.1112	0.3586	3.23	0.4698	0.76
0.9	0.5934	0.8347	0.4953	0.0981	0.3394	3.46	0.4375	0.78
1	0.6321	0.8647	0.5466	0.0855	0.3181	3.72	0.4036	0.79
1.1	0.6671	0.8892	0.5932	0.0739	0.296	4	0.3699	0.8
1.2	0.6988	0.9093	0.6354	0.0634	0.2739	4.32	0.3373	0.81
1.3	0.7275	0.9257	0.6734	0.054	0.2523	4.67	0.3063	0.82
1.4	0.7534	0.9392	0.7076	0.0458	0.2316	5.06	0.2774	0.83
1.5	0.7769	0.9502	0.7382	0.0387	0.212	5.48	0.2507	0.85
1.6	0.7981	0.9592	0.7656	0.0325	0.1937	5.95	0.2262	0.86
1.7	0.8173	0.9666	0.79	0.0273	0.1766	6.47	0.2039	0.87
1.8	0.8347	0.9727	0.8119	0.0228	0.1608	7.05	0.1836	0.88
1.9	0.8504	0.9776	0.8314	0.019	0.1462	7.69	0.1652	0.88
2	0.8647	0.9817	0.8488	0.0158	0.1329	8.39	0.1487	0.89

600

FIG. 6

m_r	Percentage of fetal DNA in the sample									
	10%		25%		50%		100%			
	Percentage of informative wells	P_r	Percentage of informative wells	P_r	Percentage of informative wells	P_r	Percentage of informative wells	P_r		
0.1	18%	0.51	18%	0.53	19%	0.56	21%	0.61		
0.2	30%	0.51	31%	0.53	32%	0.56	35%	0.61		
0.3	39%	0.51	40%	0.53	41%	0.57	43%	0.62		
0.4	45%	0.51	45%	0.54	46%	0.57	48%	0.63		
0.5	48%	0.52	49%	0.54	49%	0.57	51%	0.63		
0.6	50%	0.52	50%	0.54	50%	0.58	51%	0.64		
0.7	50%	0.52	50%	0.54	50%	0.58	50%	0.65		
0.8	49%	0.52	49%	0.54	49%	0.58	48%	0.65		
0.9	48%	0.52	47%	0.55	47%	0.59	46%	0.66		
1	46%	0.52	45%	0.55	44%	0.59	43%	0.67		
1.1	44%	0.52	43%	0.55	42%	0.6	40%	0.68		
1.2	41%	0.52	40%	0.55	39%	0.6	37%	0.69		
1.3	39%	0.52	38%	0.55	36%	0.6	34%	0.69		
1.4	36%	0.52	35%	0.56	33%	0.61	31%	0.7		
1.5	34%	0.52	33%	0.56	31%	0.61	28%	0.71		
1.6	31%	0.52	30%	0.56	28%	0.62	26%	0.72		
1.7	29%	0.53	28%	0.56	26%	0.62	23%	0.73		
1.8	27%	0.53	25%	0.57	24%	0.63	21%	0.73		
1.9	24%	0.53	23%	0.57	21%	0.63	19%	0.74		
2	22%	0.53	21%	0.57	20%	0.64	17%	0.75		

700

FIG. 7

Proportion of informative wells containing the overrepresented allele (P_r)
0.5
0.6
0.7
0.8
0.9
1.0
800
0
100
200
300
400
No. of informative wells
m_r = 0.1
m_r = 0.5
m_r = 1.0

FIG. 8

m_r	Classification	New algorithm		Old algorithm		m_r	Classification	New algorithm		Old algorithm	
		Euploid	Trisomy 21	Euploid	Trisomy 21			Euploid	Trisomy 21	Euploid	Trisomy 21
	Correct	0%	37%	0%	35%		correct	89%	92%	14%	35%
	Incorrect	3%	2%	3%	1%		Incorrect	1%	0%	4%	0%
0.1	Unclassified	97%	61%	97%	63%	1.1	Unclassified	10%	7%	81%	65%
	Correct	6%	64%	0%	40%		correct	90%	92%	11%	35%
	Incorrect	4%	2%	4%	0%		Incorrect	1%	0%	4%	0%
0.2	Unclassified	90%	34%	96%	60%	1.2	Unclassified	9%	8%	84%	65%
	Correct	34%	75%	6%	39%		correct	91%	92%	7%	34%
	Incorrect	4%	2%	4%	0%		Incorrect	1%	1%	4%	0%
0.3	Unclassified	62%	24%	90%	61%	1.3	Unclassified	8%	7%	89%	66%
	Correct	54%	81%	14%	39%		correct	92%	91%	5%	34%
	Incorrect	3%	1%	4%	0%		Incorrect	1%	1%	4%	0%
0.4	Unclassified	43%	19%	82%	61%	1.4	Unclassified	8%	9%	91%	66%
	Correct	64%	85%	20%	38%		correct	92%	90%	2%	32%
	Incorrect	3%	1%	4%	0%		Incorrect	1%	1%	4%	0%
0.5	Unclassified	33%	14%	76%	62%	1.5	Unclassified	8%	9%	93%	68%
	Correct	74%	88%	23%	38%		correct	92%	89%	1%	31%
	Incorrect	2%	1%	4%	0%		Incorrect	1%	1%	4%	0%
0.6	Unclassified	24%	11%	72%	62%	1.6	Unclassified	8%	10%	94%	69%
	Correct	80%	89%	25%	37%		correct	92%	89%	1%	30%
	Incorrect	2%	1%	4%	0%		Incorrect	1%	1%	4%	0%
0.7	Unclassified	18%	10%	71%	63%	1.7	Unclassified	7%	10%	96%	70%
	Correct	83%	91%	23%	37%		correct	92%	88%	0%	28%
	Incorrect	2%	1%	4%	0%		Incorrect	1%	1%	4%	0%
0.8	Unclassified	15%	9%	73%	63%	1.8	Unclassified	7%	11%	96%	72%
	Correct	86%	91%	21%	36%		correct	92%	87%	0%	26%
	Incorrect	1%	1%	4%	0%		Incorrect	1%	1%	3%	0%
0.9	Unclassified	12%	8%	75%	64%	1.9	Unclassified	8%	12%	97%	74%
	Correct	88%	92%	19%	36%		correct	91%	86%	0%	24%
	Incorrect	1%	0%	4%	0%		Incorrect	1%	1%	4%	0%
1	Unclassified	11%	8%	77%	64%	2	Unclassified	9%	13%	96%	76%

900

FIG. 9A

m_r	Classification	New algorithm		Old algorithm		m_r	Classification	New algorithm		Old algorithm	
		Euploid	Trisomy 21	Euploid	Trisomy 21			Euploid	Trisomy 21	Euploid	Trisomy 21
	correct	57%	88%	48%	89%		correct	100%	100%	92%	90%
	Incorrect	3%	1%	3%	1%		Incorrect	0%	0%	1%	0%
0.1	Unclassified	40%	11%	49%	10%	1.1	Unclassified	0%	0%	7%	10%
	correct	89%	98%	79%	90%		correct	100%	100%	90%	90%
	Incorrect	1%	0%	2%	0%		Incorrect	0%	0%	1%	0%
0.2	Unclassified	10%	2%	19%	10%	1.2	Unclassified	0%	0%	9%	10%
	correct	96%	99%	88%	90%		correct	100%	100%	89%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	1%	0%
0.3	Unclassified	4%	1%	11%	10%	1.3	Unclassified	0%	0%	10%	10%
	correct	99%	100%	92%	90%		correct	100%	100%	87%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	1%	0%
0.4	Unclassified	1%	0%	7%	10%	1.4	Unclassified	0%	0%	12%	10%
	correct	100%	100%	94%	90%		correct	100%	100%	85%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	1%	0%
0.5	Unclassified	0%	0%	6%	10%	1.5	Unclassified	0%	0%	13%	10%
	correct	100%	100%	94%	90%		correct	100%	100%	82%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	1%	0%
0.6	Unclassified	0%	0%	5%	10%	1.6	Unclassified	0%	0%	17%	10%
	correct	100%	100%	95%	90%		correct	100%	100%	79%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	2%	0%
0.7	Unclassified	0%	0%	5%	10%	1.7	Unclassified	0%	0%	19%	10%
	correct	100%	100%	95%	90%		correct	100%	100%	74%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	2%	0%
0.8	Unclassified	0%	0%	5%	10%	1.8	Unclassified	0%	0%	24%	10%
	correct	100%	100%	94%	90%		correct	100%	100%	70%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	2%	0%
0.9	Unclassified	0%	0%	6%	10%	1.9	Unclassified	0%	0%	28%	10%
	correct	100%	100%	94%	90%		correct	100%	100%	67%	90%
	Incorrect	0%	0%	1%	0%		Incorrect	0%	0%	3%	0%
1	Unclassified	0%	0%	6%	10%	2	Unclassified	0%	0%	30%	10%

950

FIG. 9B

m_r	Classification result	*Euploid*	*T21*	m_r	Classification result	*Euploid*	*T21*
0.1	Correct	57%	88%	0.9	Correct	100%	100%
	Incorrect	3%	1%		Incorrect	0%	0%
	Unclassified	40%	11%		Unclassified	0%	0%
0.2	Correct	89%	98%	1	Correct	100%	100%
	Incorrect	1%	0%		Incorrect	0%	0%
	Unclassified	10%	2%		Unclassified	0%	0%
0.3	Correct	96%	99%	1.1	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	4%	1%		Unclassified	0%	0%
0.4	Correct	99%	100%	1.2	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	1%	0%		Unclassified	0%	0%
0.5	Correct	100%	100%	1.3	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	0%	0%		Unclassified	0%	0%
0.6	Correct	100%	100%	1.4	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	0%	0%		Unclassified	0%	0%
0.7	Correct	100%	100%	1.5	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	0%	0%		Unclassified	0%	0%
0.8	Correct	100%	100%	2	Correct	100%	100%
	Incorrect	0%	0%		Incorrect	0%	0%
	Unclassified	0%	0%		Unclassified	0%	0%

1000

FIG. 10

m_r	Classification result	*Euploid*	*Trisomy 21*	m_r	Classification result	*Euploid*	*Trisomy 21*
	Correct	38%	44%		Correct	95%	94%
	Incorrect	2%	2%		Incorrect	0%	1%
0.1	Unclassified	60%	54%	**0.7**	Unclassified	5%	5%
	Correct	64%	69%		Correct	95%	95%
	Incorrect	2%	2%		Incorrect	1%	0%
0.2	Unclassified	34%	29%	**0.8**	Unclassified	4%	5%
	Correct	78%	82%		Correct	96%	96%
	Incorrect	2%	1%		Incorrect	0%	0%
0.3	Unclassified	20%	17%	**0.9**	Unclassified	4%	4%
	Correct	85%	87%		Correct	97%	96%
	Incorrect	1%	1%		Incorrect	0%	0%
0.4	Unclassified	14%	12%	**1**	Unclassified	3%	4%
	Correct	89%	90%		Correct	97%	96%
	Incorrect	1%	1%		Incorrect	0%	0%
0.5	Unclassified	10%	9%	**1.5**	Unclassified	3%	4%
	Correct	92%	93%		Correct	96%	95%
	Incorrect	1%	1%		Incorrect	0%	0%
0.6	Unclassified	7%	7%	**2**	Unclassified	4%	5%

1100

FIG. 11

Number of wells	Classification results	Fetal DNA percentage in the sample							
		10%		25%		50%		100%	
		Euploid	*Trisomy 21*	*Euploid*	*Trisomy 21*	*Euploid*	*Trisomy 21*	*Euploid*	*Trisomy 21*
384	Correct	0%	0%	6%	7%	47%	49%	89%	90%
	Incorrect	0%	0%	0%	1%	2%	2%	1%	1%
	Unclassified	100%	100%	93%	92%	51%	49%	10%	9%
768	Correct	0%	0%	23%	24%	73%	76%	98%	99%
	Incorrect	0%	0%	2%	2%	2%	1%	0%	0%
	Unclassified	100%	100%	75%	74%	25%	23%	2%	1%
1152	Correct	1%	1%	39%	40%	86%	87%	100%	100%
	Incorrect	0%	0%	2%	2%	1%	1%	0%	0%
	Unclassified	99%	99%	59%	58%	13%	12%	0%	0%
1536	Correct	2%	2%	50%	51%	93%	93%	100%	100%
	Incorrect	0%	0%	2%	2%	1%	1%	0%	0%
	Unclassified	98%	97%	48%	47%	7%	6%	0%	0%
1920	Correct	4%	4%	60%	60%	96%	96%	100%	100%
	Incorrect	0%	1%	2%	2%	0%	0%	0%	0%
	Unclassified	96%	95%	38%	37%	4%	3%	0%	0%
3840	Correct	18%	19%	84%	83%	100%	100%	100%	100%
	Incorrect	1%	1%	1%	1%	0%	0%	0%	0%
	Unclassified	81%	80%	15%	15%	18%	22%	0%	0%
7680	Correct	43%	45%	97%	97%	100%	100%	100%	100%
	Incorrect	2%	2%	0%	0%	0%	0%	0%	0%
	Unclassified	55%	53%	3%	3%	0%	0%	0%	0%

1200

FIG. 12

sample	geno-type[a]	no. of wells	no. of wells positive for individual alleles				m_r[b]	SPRT result		
			A only	G only	AG	all negative		P_r[c]	unclassifiable region[d]	classification[e]
Placental DNA										
N677	AG	384	85	83	126	90	0.79	0.51	0.63 - 0.65	euploid
N710	AG	384	102	83	73	126	0.52	0.55	0.61 - 0.63	euploid
N435	AGG	384	49	157	130	48	0.63	0.76	0.62 - 0.64	T21
N981	AAG	384	135	69	82	98	0.5	0.66	0.61 - 0.63	T21
Placental RNA										
V533	AG	384	103	93	71	117	0.56	0.53	0.61 - 0.63	euploid
V943	AG	384	89	100	74	121	0.55	0.53	0.61 - 0.63	euploid
N435	AGG	384	52	138	95	99	0.48	0.73	0.61 - 0.63	T21
T215	AAG	384	146	58	138	42	0.71	0.72	0.62 - 0.64	T21

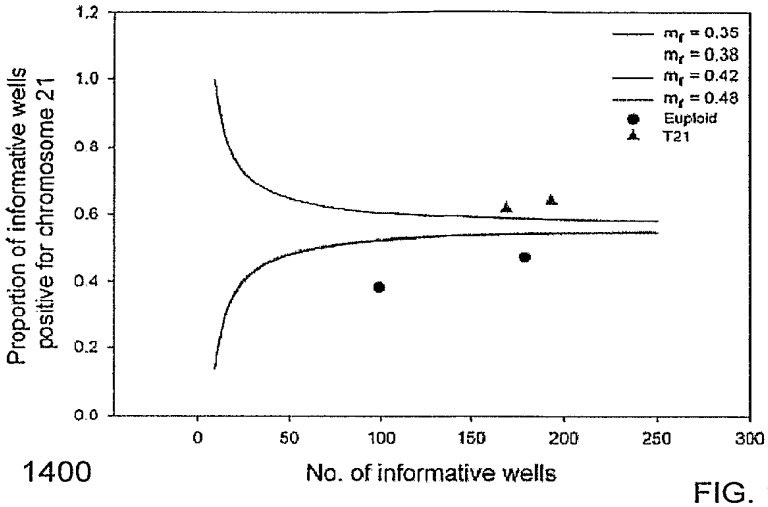
1300

FIG. 13A

sample	geno-type	no. of wells	no. of wells positive for individual alleles				m	SPRT result		
			A only	G only	AG	all negative		ratio	unclassifiable region	classification
M2390P	AG	384	90	100	97	97	0.67	0.526	0.62 - 0.64	euploid
M2391P	AG	384	97	105	65	117	0.55	0.52	0.61 - 0.63	euploid
M2473P	AG	384	66	92	34	192	0.3	0.582	0.59 - 0.62	euploid
M2524P	AG	384	29	28	3	324	0.08	0.509	0.54 - 0.64	euploid
M2528P	AG	384	112	85	44	143	0.41	0.569	0.60 - 0.62	euploid
M2601P	AG	384	90	101	72	121	0.55	0.529	0.61 - 0.63	euploid
M2607P	AG	384	73	91	57	163	0.41	0.555	0.60 - 0.63	euploid
M2638P	AG	384	66	90	52	176	0.37	0.577	0.59 - 0.62	euploid
M2639P	AG	384	71	56	17	240	0.21	0.559	0.58 - 0.62	euploid
M2525P	AAG	384	110	53	21	200	0.21	0.675	0.58 - 0.61	T21
M2272P	AAG	768	246	127	112	283	0.37	0.66	0.60 - 0.61	T21
M2718P	AGG	384	66	114	66	138	0.42	0.633	0.60 - 0.62	T21
M1519P	AGG	384	58	130	54	142	0.34	0.691	0.59 - 0.62	T21

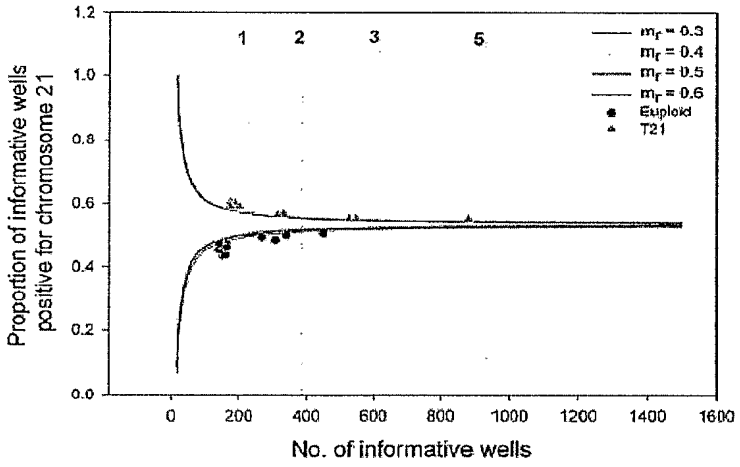
1350

FIG. 13B



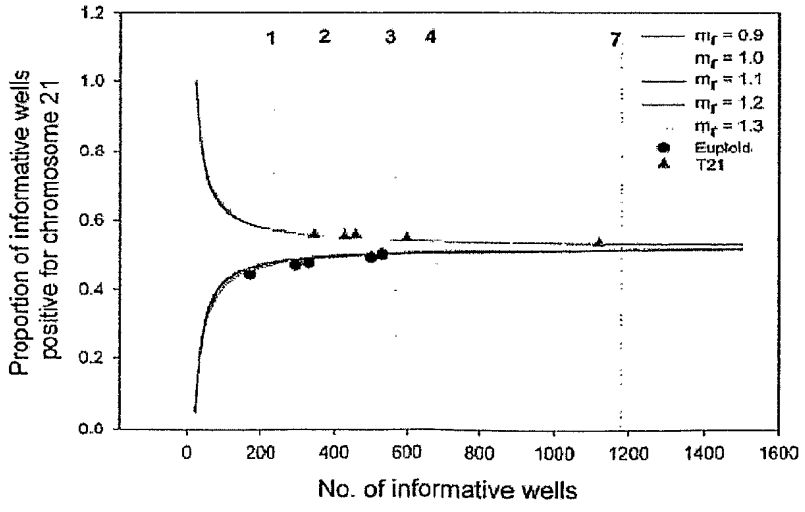
1400

FIG. 14A



1430

FIG. 14B



1460

FIG. 14C

sample	genotype	no. of wells	no. of wells positive for individual alleles				m_r[a]	SPRT result		
			A only	G only	GA	all negative		P_r[b]	unclassifiable region	classification[c]
V943	AG	765	168	209	117	271	0.47	0.554	0.609- 0.621	euploid
N435	AGG	765	92	281	105	287	0.30	0.753	0.597- 0.610	T21
T21-5	AAG	765	267	126	141	231	0.43	0.679	0.607- 0.618	T21

[a] no. of reference molecules per well

[b] P_r values were calculated by the equation:

$$\frac{\text{no. of wells positive for the overrepresented allele}}{\text{no. of wells positive for A only + no of wells positive for G only}}$$

[c] "Euploid" was assigned when the Pr was below the unclassifiable region; "T21" was assigned when the Pr was abov e the unclassifiable region. T21, trisomy 21

FIG. 15A

panel #	no. of wells	no. wells positive for individual alleles			
		A only	G only	GA	all negative
S1	765	14	15	0	736
S2	765	7	10	0	748
S3	765	13	9	1	744
S4	765	7	8	0	750
S5	765	12	8	0	745
S6	765	8	14	0	743
S7	765	13	9	0	743
S8	765	9	8	0	748
S9	765	14	15	0	736
S10	765	9	12	0	744
S11	765	8	17	0	740
S12	765	11	9	0	745

FIG. 15B

sample	geno-type	panels[a]	no. of wells positive for individual alleles					SPRT result		
			no of wells	A only	G only	GA	all negative	P_r	unclassifiable region	classification
M2879P	AG	5	3825	82	103	5	3635	0.557	0.570- 0.602	Euploid
M2875P	AG	3	2295	35	40	0	2220	0.533	0.546- 0.625	Euploid
M3078P	AG	3	2295	33	33	1	2228	0.500	0.541- 0.631	Euploid
M2976P	AG	12	9180	56	46	0	9078	0.549	0.556- 0.615	Euploid
M2831P	AAG	2	1530	42	13	0	1475	0.764	0.531- 0.640	T21

[a] no. of panels needed for classification

FIG. 15C

Sample	Genotype[a]	No. of wells positive for A or G allele			m_r[b]	SPRT analysis		
		A only	G only	Both		P_r[c]	Unclassifiable region[d]	Classification[e]
N3505	AG	157	131	48	0.27	0.545	0.59-0.61	Euploid
N3530	AG	174	132	52	0.28	0.569	0.59-0.61	Euploid
N3531	AG	146	129	28	0.23	0.531	0.58-0.60	Euploid
N3533	AG	149	135	46	0.27	0.525	0.59-0.61	Euploid
N2772	AG	159	149	79	0.35	0.516	0.60-0.61	Euploid
N3506	AG	154	145	66	0.32	0.515	0.60-0.61	Euploid
N3555	AG	94	73	7	0.11	0.563	0.58-0.61	Euploid
N3560	AG	101	109	28	0.18	0.519	0.58-0.61	Euploid
N3561	AG	91	70	17	0.12	0.565	0.58-0.61	Euploid
N828	AAG	153	89	22	0.16	0.632	0.58-0.60	T18
CA2	AGG	12	28	0	0.016	0.700	0.51-0.66	T18
CA7	AGG	97	171	56	0.22	0.638	0.59-0.61	T18

[a] Genotypes were determined by mass spectrometry.

[b] The m_r value indicates the average number of reference template molecules per well.

[c] The P_r value indicates the proportion of informative wells positive for the overrepresented allele.

[d] The unclassifiable region varies with the m_r value.

[e] "Euploid" was assigned when P_r was below the unclassifiable region; "T18" was assigned when P_r was above the unclassifiable region.

FIG. 16A

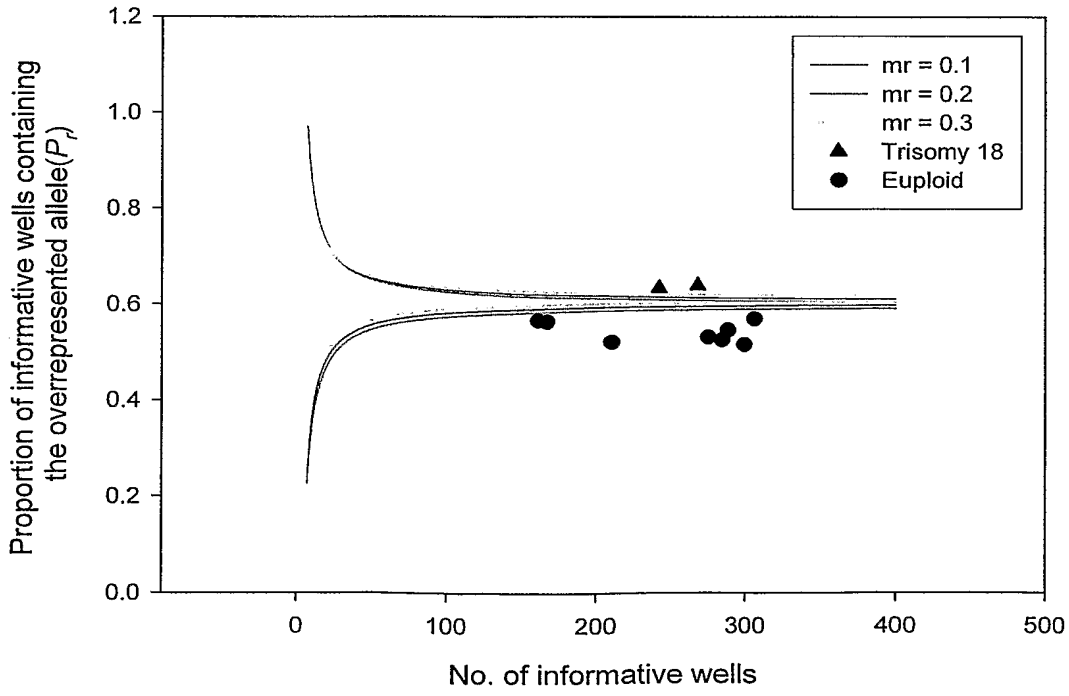


FIG. 16B

Type	Cases	Total No. of panels used	Total No. of informative wells	No. of well count			m_r[a]	SPRT analysis		
				Both positive	Chr21 only	Chr1 only		P_r[b]	Unclassifiable region	Classification[c]
T21	N0234	1	303	68	172	131	0.30	0.568	0.507 - 0.559	T21
	N0274	2	558	118	307	251	0.28	0.550	0.518 - 0.547	T21
	N0332	1	293	73	168	125	0.30	0.573	0.506 - 0.560	T21
	N0349	2	574	107	322	252	0.27	0.561	0.518 - 0.546	T21
	N0435	1	299	67	178	123	0.29	0.589	0.506 - 0.559	T21
	N0456	1	318	68	189	129	0.30	0.594	0.508 - 0.558	T21
	N0667	2	602	134	345	257	0.30	0.573	0.519 - 0.546	T21
	N0764	1	289	62	167	122	0.28	0.578	0.505 - 0.560	T21
	N0890	4	1206	272	657	549	0.31	0.545	0.526 - 0.539	T21
	N0891	1	317	103	180	137	0.38	0.568	0.510 - 0.558	T21
Normal	N0196	1	262	68	111	151	0.34	0.424	0.504 - 0.563	Euploid
	N0198	1	285	68	132	153	0.34	0.463	0.506 - 0.561	Euploid
	N0230	2	627	109	272	255	0.27	0.516	0.517 - 0.548	Euploid
	N0231	2	542	98	275	267	0.27	0.507	0.517 - 0.547	Euploid
	N0262	1	295	57	144	151	0.32	0.488	0.506 - 0.560	Euploid
	N0463	1	263	70	120	143	0.33	0.456	0.504 - 0.563	Euploid
	N0527	1	292	56	137	155	0.32	0.469	0.506 - 0.560	Euploid
	N0619	1	281	60	138	143	0.31	0.491	0.505 - 0.561	Euploid
	N0625	1	291	60	134	157	0.33	0.460	0.507 - 0.560	Euploid
	N0651	2	566	118	287	279	0.30	0.507	0.519 - 0.547	Euploid

[a] no. of reference molecules per well

[b] P_r values were calcualated by the equation:

$$\frac{\text{no. of wells positive for the overrepresented allele (chr21)}}{\text{no. of wells positive for chr21 only + no of wells positive for chr1 only}}$$

[c] "Euploid" was assigned when the Pr was below the unclassifiable region; "T21" was assigned when the Pr was above the unclassifiable region.

FIG. 17

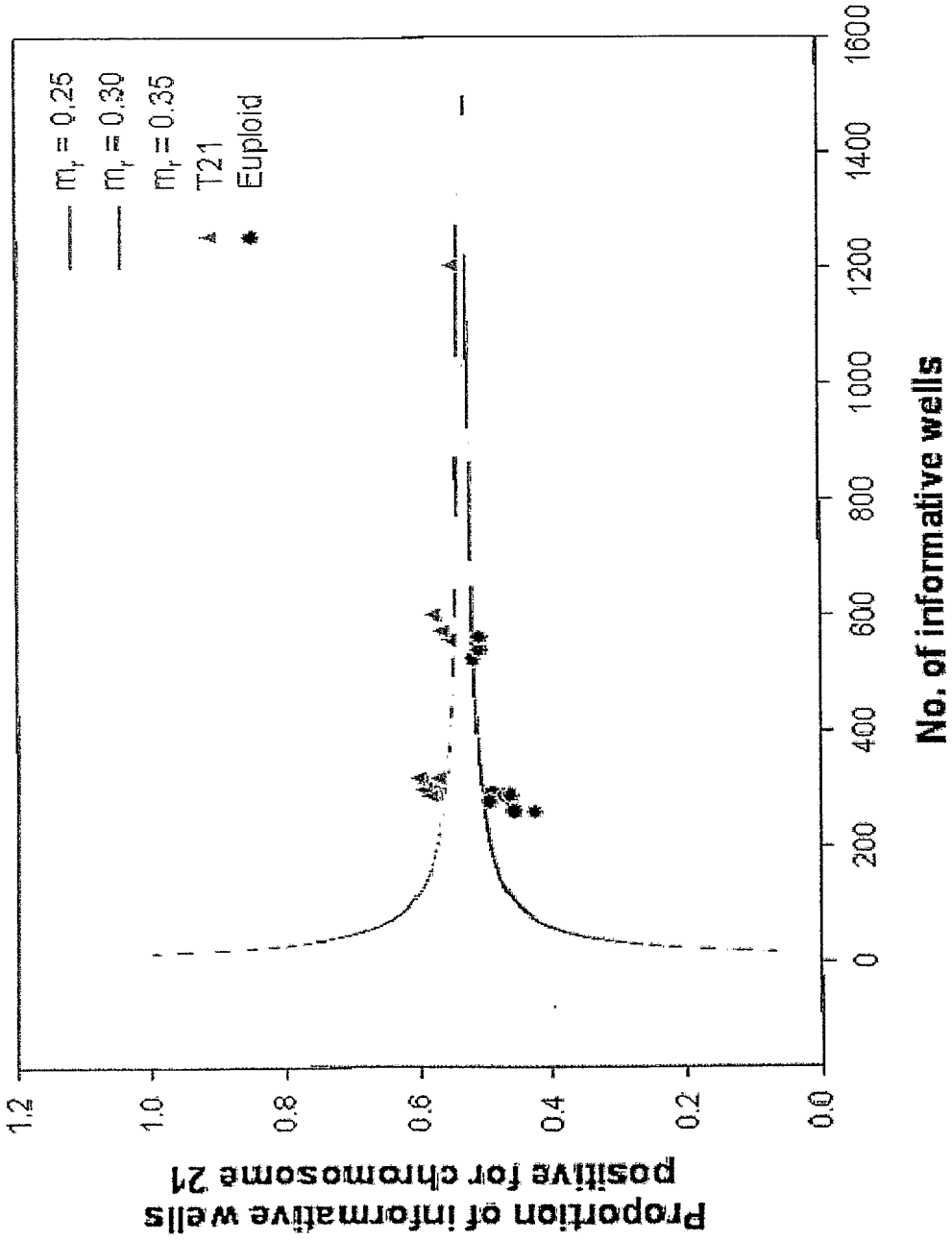

Proportion of informative wells positive for chromosome 21
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
1200
1400
1600
No. of informative wells
m_r = 0.25
m_r = 0.30
m_r = 0.35
T21
Euploid

FIG. 18

Type	Cases	No. of panels used	No. of informative wells	No. of well count			SPRT analysis			
				Both positive	Chr21 only	Chr1 only	m_r[a]	P_r[b]	Unclassifiable region	Classification[c]
T21	A0274	1	380	269	249	131	0.74	0.865	0.567 - 0.585	T21
	A2008	1	348	373	233	115	1.02	0.870	0.579 - 0.596	T21
	A2178	1	384	288	239	145	0.83	0.822	0.571 - 0.588	T21
	AS383A	1	364	109	244	120	0.38	0.870	0.561 - 0.573	T21
	AS979A	1	258	492	188	88	1.32	0.734	0.590 - 0.610	T21
	AS3457A	1	378	83	256	120	0.31	0.881	0.549 - 0.572	T21
Normal	A0238	1	387	275	185	182	0.84	0.504	0.578 - 0.582	Euploid
	A0237	1	392	228	184	208	0.84	0.489	0.572 - 0.588	Euploid
	A0255	1	378	222	185	183	0.75	0.516	0.568 - 0.585	Euploid
	A0283	1	254	43	127	127	0.25	0.500	0.541 - 0.578	Euploid
	AS365A	1	370	135	192	178	0.53	0.518	0.558 - 0.578	Euploid
	AS483A	1	354	84	188	188	0.38	0.525	0.550 - 0.574	Euploid
	AS484A	1	372	271	194	178	0.88	0.522	0.573 - 0.580	Euploid
	AS780A	1	413	188	200	213	0.74	0.484	0.588 - 0.584	Euploid
	AS780A	1	316	67	159	157	0.35	0.503	0.548 - 0.575	Euploid
	AS897A	1	198	8	87	111	0.17	0.439	0.532 - 0.579	Euploid
	AS921A	1	152	8	89	83	0.13	0.454	0.523 - 0.586	Euploid
	AS980A	1	394	137	191	203	0.59	0.485	0.581 - 0.580	Euploid
	AS2282A	1	334	342	177	157	1.06	0.530	0.580 - 0.587	Euploid
	AS2425A	1	128	8	84	84	0.10	0.500	0.518 - 0.591	Euploid
	AS2434A	1	379	97	203	176	0.44	0.536	0.555 - 0.575	Euploid
	A1119	1	354	123	185	189	0.48	0.523	0.555 - 0.577	Euploid
	A1122	1	369	100	195	174	0.44	0.528	0.554 - 0.576	Euploid
	A1152	1	258	48	128	128	0.28	0.500	0.542 - 0.578	Euploid
	A1180	1	246	498	133	113	1.61	0.641	0.604 - 0.622	Euploid
	A1275	1	373	95	201	172	0.43	0.538	0.554 - 0.575	Euploid
	A1375	1	369	55	190	169	0.35	0.528	0.550 - 0.573	Euploid
	A1388	1	367	87	165	182	0.47	0.462	0.565 - 0.577	Euploid
	A1399	1	371	114	197	174	0.47	0.531	0.556 - 0.577	Euploid

[a] no. of reference molecules per well

[b] P_r values were calculated by the equation:

$$\frac{\text{no. of wells positive for the overrepresented allele (chr21)}}{\text{no. of wells positive for chr21 only + no of wells positive for chr1 only}}$$

[c] "Euploid" was assigned when the Pr was below the unclassifiable region; "T21" was assigned when the Pr was above the unclassifiable region.

FIG. 19

Type	Cases	Total No. of panels used	Total No. of informative wells	No. of well count			m_r [a]	SPRT analysis		
				Both positive	Chr18 only	Chr21 only		P_r [b]	Unclassifiable region	Classification [c]
Euploid	N0230	1	380	243	181	199	0.86	0.476	0 .573 - 0.589	E
	N0625	1	370	225	179	191	0.78	0.484	0 .569 - 0.586	E
	N0651	1	358	206	192	166	0.67	0.536	0.564 - 0.583	E
T18	N0828	1	543	256	329	214	0.95	0.606	0.579 - 0.590	T18
	N0830	1	346	353	240	106	0.92	0.694	0.574 - 0.592	T18
	N0936	1	372	285	209	163	0.88	0.562	0.573 - 0.590	E
	N1678	1	343	115	221	122	0.37	0.644	0.551 - 0.574	T18
	N1701	1	303	94	191	112	0.31	0.630	0.546 - 0.575	T18

[a] no. of reference molecules per well

[b] P_r values were calcualated by the equation:

$$\frac{\text{no. of wells positive for the overrepresented allele (chr18)}}{\text{no. of wells positive for chr21 only + no of wells positive for chr18 only}}$$

[c] "Euploid" was assigned when the Pr was below the unclassifiable region; "T18" was assigned when the Pr was above the unclassifiable region.

FIG. 20

Type	Cases	Plate No.	Assay	Total No. of informative wells	No. of well count			m_r[a]	SPRT analysis		
					Both positive	Chr 21 only	Ref only		P_r[b]	Unclassifiable region	Classification[c]
Normal	N0196	1	chr21/chr1	148	78	71	77	0.52	0.480	0.489 - 0.584	E
		1	chr21/chr18	152	63	79	73	0.44	0.520	0.487 - 0.584	U
			SUM	300	141	150	150	0.48	0.500	0.512 - 0.560	E
	N0198	1	chr21/chr1	162	73	70	92	0.56	0.432	0.490 - 0.581	E
		1	chr21/chr18	183	68	93	90	0.53	0.508	0.498 - 0.575	U
			SUM	345	141	163	182	0.55	0.472	0.517 - 0.557	E
	N0230	1	chr21/chr1	162	58	81	81	0.45	0.500	0.490 - 0.581	U
		1	chr21/chr18	165	45	81	84	0.41	0.491	0.489 - 0.580	U
			SUM	327	103	162	165	0.43	0.495	0.512 - 0.558	E
	N0231	1	chr21/chr1	173	53	81	92	0.47	0.468	0.494 - 0.578	E
		1	chr21/chr18	166	51	85	81	0.42	0.512	0.490 - 0.580	U
			SUM	339	104	166	173	0.45	0.490	0.514 - 0.557	E
	N0262	1	chr21/chr1	186	97	101	85	0.64	0.543	0.503 - 0.575	U
		1	chr21/chr18	176	111	91	85	0.71	0.517	0.504 - 0.557	U
			SUM	362	208	192	170	0.68	0.530	0.522 - 0.557	U
		2	chr21/chr1	174	103	85	89	0.69	0.489	0.503 - 0.577	E
			SUM	536	311	277	259	0.68	0.517	0 .523 - 0.552	E
T21	N0234	1	chr21/chr1	193	68	108	85	0.51	0.560	0.500 - 0.573	U
		1	chr21/chr18	178	65	92	86	0.50	0.517	0.496 - 0.576	U
			SUM	371	133	200	171	0.50	0.539	0.517 - 0.556	U
		2	chr21/chr1	178	73	92	86	0.53	0.517	0.498 - 0.576	U
			SUM	549	206	292	257	0.51	0.532	0 .524 - 0.549	U
		2	chr21/chr18	178	62	95	83	0.47	0.534	0.495 - 0.576	U
			SUM	727	268	387	340	0.50	0.532	0 .527 - 0.546	U
		3	chr21/chr1	184	108	103	81	0.68	0.560	0.504 - 0.575	U
			SUM	911	376	490	421	0.54	0.538	0.529 - 0.545	U
		3	chr21/chr18	198	102	120	78	0.63	0.606	0.505 - 0.572	T21
			SUM	1109	478	610	499	0.55	0.550	0.531 - 0.544	T21
	N0332	1	chr21/chr1	173	109	103	70	0.63	0.595	0.500 - 0.577	T21
		1	chr21/chr18	191	82	110	81	0.55	0.576	0.501 - 0.574	T21
			SUM	364	191	213	151	0.59	0.585	0.519 - 0.557	T21
	N0349	1	chr21/chr1	158	62	83	75	0.44	0.525	0.489 - 0.582	U
		1	chr21/chr18	166	46	87	79	0.39	0.524	0.489 - 0.580	U
			SUM	324	108	170	154	0.42	0.525	0.512 - 0.558	U
		2	chr21/chr1	149	66	86	63	0.41	0.577	0.485 - 0.585	U
			SUM	473	174	256	217	0.41	0.541	0.519 - 0.551	U
		2	chr21/chr18	168	43	100	68	0.34	0.595	0.487 - 0.580	T21
			SUM	641	217	356	285	0.40	0.555	0.523 - 0.546	T21
	N0764	1	chr21/chr1	182	82	106	76	0.53	0.582	0.498 - 0.575	T21
		1	chr21/chr18	178	83	97	81	0.56	0.545	0.498 - 0.576	U
			SUM	360	165	203	157	0.54	0.564	0.518 - 0.556	T21
	N0890	1	chr21/chr1	180	103	108	72	0.61	0.600	0.501 - 0.576	T21
		1	chr21/chr18	188	68	104	84	0.50	0.553	0.499 - 0.574	U
			SUM	368	171	212	156	0.55	0.576	0.518 - 0.556	T21

[a] no. of reference molecules per well

[b] P_r values were calcualated by the equation:

$$\frac{\text{no. of wells positive for the overrepresented allele (chr21)}}{\text{no. of wells positive for chr21 only + no of wells positive for Ref only}}$$

[c] "Euploid" was assigned when the Pr was below the unclassifiable region; "T21" was assigned when the Pr was above the unclassifiable region.

FIG. 21

			Poisson corrected total no. Chromosome 21	Reference chr	P_r	Classification
Euploid	Case 1	Assay 1	37	33	0.52345751	**Unclass**
		Assay 2	31	42	0.42506276	**Unclass**
		Assay 3	29	37	0.44198785	**Unclass**
		Assay 4	39	26	0.6008691	**Unclass**
		Plate 1 subtotal	136	138	0.495821	**Unclass**
		Assay 1	32	25	0.56605316	**Unclass**
		Assay 2	27	24	0.53165424	**Unclass**
		Assay 3	10	17	0.3681535	**Unclass**
		Assay 4	24	18	0.5626764	**Unclass**
		Plate 2 subtotal	93	84	0.52474	**Unclass**
		Total	229	223	0.507182	**Euploid**
	Case 2	Assay 1	45	35	0.56321585	**Unclass**
		Assay 2	31	45	0.41245525	**Unclass**
		Assay 3	32	22	0.59954763	**Unclass**
		Assay 4	38	35	0.5228175	**Unclass**
		Plate 1 subtotal	146	135	0.519153	**Unclass**
		Assay 1	28	23	0.55272135	**Unclass**
		Assay 2	23	22	0.51196923	**Unclass**
		Assay 3	23	15	0.59688187	**Unclass**
		Assay 4	29	26	0.5293349	**Unclass**
		Plate 2 subtotal	102	85	0.545204	**Unclass**
		Plate 1+2 subtotal	248	221	0.529583	**Unclass**
		Assay 1	17	25	0.4122149	**Unclass**
		Assay 2	25	16	0. 6027299	**Unclass**
		Assay 3	13	26	0.3382802	**Unclass**
		Assay 4	27	24	0.5316872	**Unclass**
		Plate 3 subtotal	82	91	0.475775	**Unclass**
		Plate 1- 3 subtotal	331	311	0.515069	**Unclass**
		Assay 1	15	9	0.6269193	**Unclass**
		Assay 2	14	15	0.4824155	**Unclass**
		Assay 3	18	17	0.5146313	**Unclass**
		Assay 4	21	18	0.5270080	**Unclass**
		Plate 4 subtotal	69	60	0.532264	**Unclass**
		Plate 1- 4 subtotal	399	372	0.517942	**Unclass**

FIG. 22A

	Assay 1	10	13	0.4337957	**Unclass**
	Assay 2	11	14	0.4390082	**Unclass**
	Assay 3	7	16	0.3018022	**Unclass**
	Assay 4	19	10	0.6579085	**Unclass**
	Plate 5 subtotal	48	54	0.469951	**Unclass**
	Total	447	426	0.5123444	**Euploid**
T21					
Case 3	Assay 1	50	33	0.60012253	**Unclass**
	Assay 2	56	47	0.54458055	**Unclass**
	Assay 3	46	28	0.6196626	**Unclass**
	Assay 4	56	36	0.6108731	**Unclass**
	Total	208	144	0.59079	**T21**
Case 4	Assay 1	56	36	0.61085513	**Unclass**
	Assay 2	38	32	0.53906912	**Unclass**
	Assay 3	40	24	0.62788068	**Unclass**
	Assay 4	47	27	0.634511	**Unclass**
	Total	180	119	0.603476	**T21**

FIG. 22B

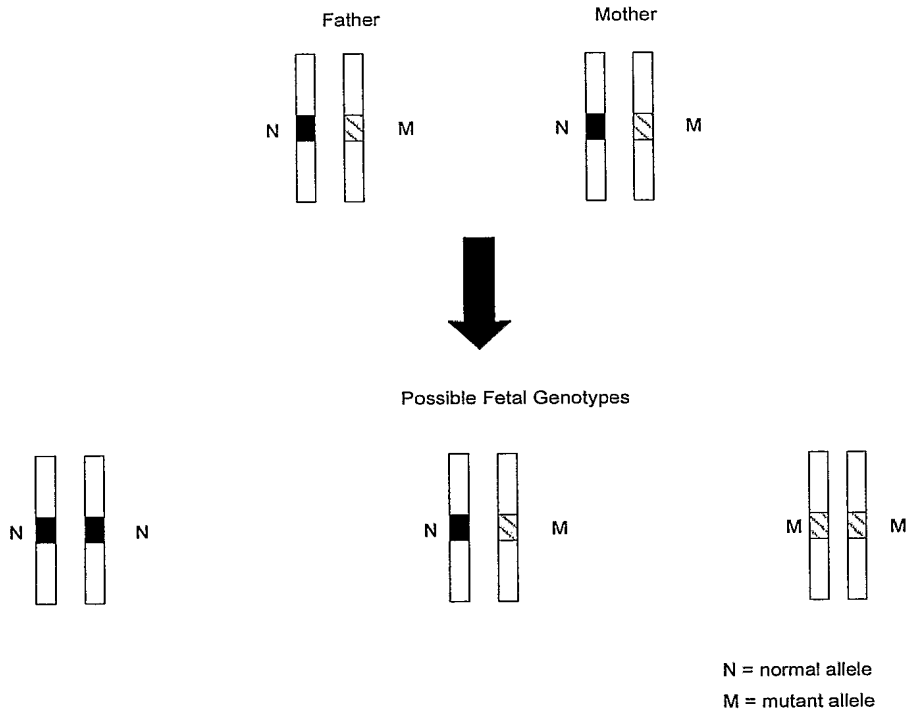

Father
Mother
N
M
N
M
Possible Fetal Genotypes
N
N
N
M
M
M
N = normal allele
M = mutant allele

FIG. 23

Cases	Samples	Expected %	No of panels	*ZFY*	*ZFX*	ALL POS	NET *ZFY*	NET *ZFX*	Ref allele	m_r	SPRT[a]
1	XY + XX	50%	1	239	465	155	84	310	Y	0.37	XX
	(XX %)	25%	1	328	447	205	123	242	Y	0.56	XX
2	$XY_1 + XY_2$	50%	1	414	405	220	194	185	X	0.85	XY
	(XY_2 %)	25%	1	432	426	245	187	181	X	0.81	XY

FIG. 24A

Cases	No of panels	*ZFY*	*ZFX*	ALL POS	NET *ZFY*	NET *ZFX*	Ref allele	m_r	SPRT
1	1	81	171	18	63	153	Y	0.11	XX
2	1	103	175	23	80	152	Y	0.14	XX
3	1	110	175	28	82	147	Y	0.16	XX
4	1	111	156	24	87	132	Y	0.16	XX
5	2	236	322	50	186	272	Y	0.17	XX
6	1	116	184	29	87	155	Y	0.16	XX
7	1	94	169	19	75	150	Y	0.13	XX
8	1	91	155	24	67	131	Y	0.13	XX
9	3	333	447	74	259	373	Y	0.16	XX
10	1	110	154	24	86	130	Y	0.16	XX
11	1	83	170	26	57	144	Y	0.11	XX
12	1	85	159	16	69	143	Y	0.12	XX

FIG. 24B

Cases	No of panels	Maternal genotype[a]	Fetal genotype[a]	Expected fetal %	A	G	ALL POS	NET A	NET G	Actual fetal %[b]	Ref allele	m_t	SPRT[d]
1	1	AG	GG	15%	264	318	116	148	202	9.6	A	0.42	GG
2	1	AG	GG	15%	208	276	81	127	195	14.3	A	0.32	GG
3	1	AG	AG	15%	248	251	87	161	164	15.4	A	0.39	AG
4	1	AG	GG	15%	158	240	51	107	189	13.7	A	0.23	GG
5	1	AG	AG	15%	226	245	70	156	175	16.8	A	0.35	AG
6	1	AG	AG	15%	319	319	130	189	189	18.1	A	0.54	AG
7	1	AG	AG	15%	254	279	94	160	185	19.8	A	0.40	AG
8	1	AG	GG	15%	296	363	144	152	219	15.7	A	0.49	GG
9	1	AG	AG	15%	279	285	100	179	185	13.0	A	0.45	AG
10	1	AG	AA	15%	337	278	120	217	158	NA[c]	G	0.45	AA
1	1	AG	GG	20%	248	295	96	152	199	13.0	A	0.39	GG
2	1	AG	GG	20%	191	289	79	112	210	19.5	A	0.29	GG
3	1	AG	AG	20%	248	218	75	173	143	18.2	G	0.34	U
4	1	AG	GG	20%	158	257	59	99	198	25.9	A	0.23	GG
5	1	AG	AG	20%	231	215	71	160	144	20.3	G	0.33	AG
6	1	AG	AG	20%	281	283	98	183	185	26.3	A	0.46	AG
7	1	AG	AG	20%	256	231	88	168	143	22.7	G	0.36	AG
8	1	AG	GG	20%	246	311	104	142	207	23.0	A	0.39	GG
9	1	AG	AG	20%	267	249	83	184	166	21.9	G	0.39	AG
10	1	AG	AA	20%	303	203	77	226	126	NA[c]	G	0.31	AA
1	1	AG	GG	25%	244	304	99	145	205	15.7	A	0.38	GG
2	1	AG	GG	25%	209	314	88	121	226	24.9	A	0.32	GG
3	1	AG	AG	25%	204	212	55	149	157	28.0	A	0.31	AG
4	1	AG	GG	25%	185	290	76	109	214	25.9	A	0.28	GG
5	1	AG	AG	25%	233	243	77	156	166	24.5	A	0.36	AG
6	1	AG	AG	25%	292	305	110	182	195	27.3	A	0.48	AG
7	1	AG	AG	25%	261	275	95	166	180	22.5	A	0.42	AG
8	1	AG	GG	25%	236	327	96	140	231	26.7	A	0.37	GG
9	1	AG	AG	25%	271	280	104	167	176	25.2	A	0.44	AG
10	1	AG	AA	25%	243	154	46	197	108	NA[c]	G	0.22	AA
1	1	AG	GG	50%	136	293	57	79	236	33.6	A	0.20	GG
2	1	AG	GG	50%	163	358	78	85	280	45.8	A	0.24	GG
3	1	AG	AG	50%	200	208	53	147	155	55.7	A	0.30	AG
4	1	AG	GG	50%	137	356	66	71	290	53.0	A	0.18	GG
5	1	AG	AG	50%	247	246	91	156	155	57.7	G	0.39	AG
6	1	AG	AG	50%	218	245	68	150	177	53.8	A	0.34	AG
7	1	AG	AG	50%	255	282	103	152	179	51.3	A	0.41	AG
8	1	AG	GG	50%	158	398	83	75	315	57.6	A	0.23	GG
9	1	AG	AG	50%	235	225	74	161	151	55.7	G	0.35	AG
10	1	AG	AA	50%	317	106	35	282	71	NA[c]	G	0.15	AA

[a] A = Mutated allele, G = Wildtype allele.
[b] The actual fetal DNA % is determined by the digital *ZFY/X* assay.
[c] The fetus is female and digital *ZFY/X* assay cannot be applied.
[d] SPRT was done with the fetal % determined by the digital *ZFY/X* assay except case 10; U, unclassified.

FIG. 25

Case	Maternal genotype[a]	Fetal genotype[a]	No of panels	M	N	ALL POS	NET M	NET N	1[b] Expected fetal %	1[b] SPRT	2[c] Fetal % by digital *ZFY/X*	2[c] SPRT	Ref Allele	m_r
1	MN	MM	3	846	763	273	573	490	5%	MM	7.8%	MM	N	0.40
	MN	MM	3	863	748	289	574	459	10%	MM	9.6%	MM	N	0.39
	MN	MM	1	323	219	79	244	140	25%	MM	24.1%	MM	N	0.34
	MN	MM	1	418	160	92	326	68	50%	MM	57.2%	MM	N	0.23
2	MN	MN	9	2559	2476	887	1672	1589	5%	U	3.4%	U	N	0.45
	MN	MN	3	895	907	360	535	547	10%	MN	9.7%	MN	M	0.49
	MN	MN	1	277	263	97	180	166	25%	MN	24.2%	MN	N	0.42
	MN	MN	1	273	295	101	172	194	50%	MN	58.0%	MN	M	0.44

[a] M = with CTTT deletion, N = without CTTT deletion.

[b] SPRT was done with the expected %.

[c] SPRT was done with the fetal % determined by the digital *ZFY/X* assay; U, unclassified

FIG. 26A

Cases	No of panels	Maternal genotype[a]	Fetal genotype[a]	Expected fetal %	M	N	ALL POS	NET M	NET N	Actual fetal %[b]	Ref allele	m_r	SPRT[c]
1	1	MN	MN	20%	203	200	58	145	142	29.0	N	0.30	MN
2	1	MN	NN	20%	149	204	39	110	165	29.7	M	0.22	NN
3	1	MN	NN	20%	128	212	31	97	181	30.7	M	0.18	NN
4	1	MN	MN	20%	173	196	45	128	151	27.0	M	0.26	MN
5	1	MN	NN	20%	147	230	47	100	183	24.3	M	0.21	NN
6	1	MN	MN	20%	199	206	45	154	161	19.2	M	0.30	MN
7	1	MN	NN	20%	165	237	58	107	179	18.7	M	0.24	NN
8	1	MN	MN	20%	203	180	43	160	137	23.4	N	0.27	MN
9	1	MN	NN	20%	173	241	61	112	180	16.9	M	0.26	NN
10	1	MN	MN	20%	174	190	52	122	138	16.8	M	0.26	U
11	1	MN	NN	20%	167	242	49	118	193	17.7	M	0.25	NN
12	1	MN	MN	20%	213	199	47	166	152	21.6	N	0.30	MN

[a] M = with CTTT deletion, N = without CTTT deletion.

[b] The actual fetal DNA % is determined by the digital *ZFXY* assay.

[c] SPRT was done with the fetal % determined by the digital *ZFY/X* assay; U, unclassified.

FIG. 26B

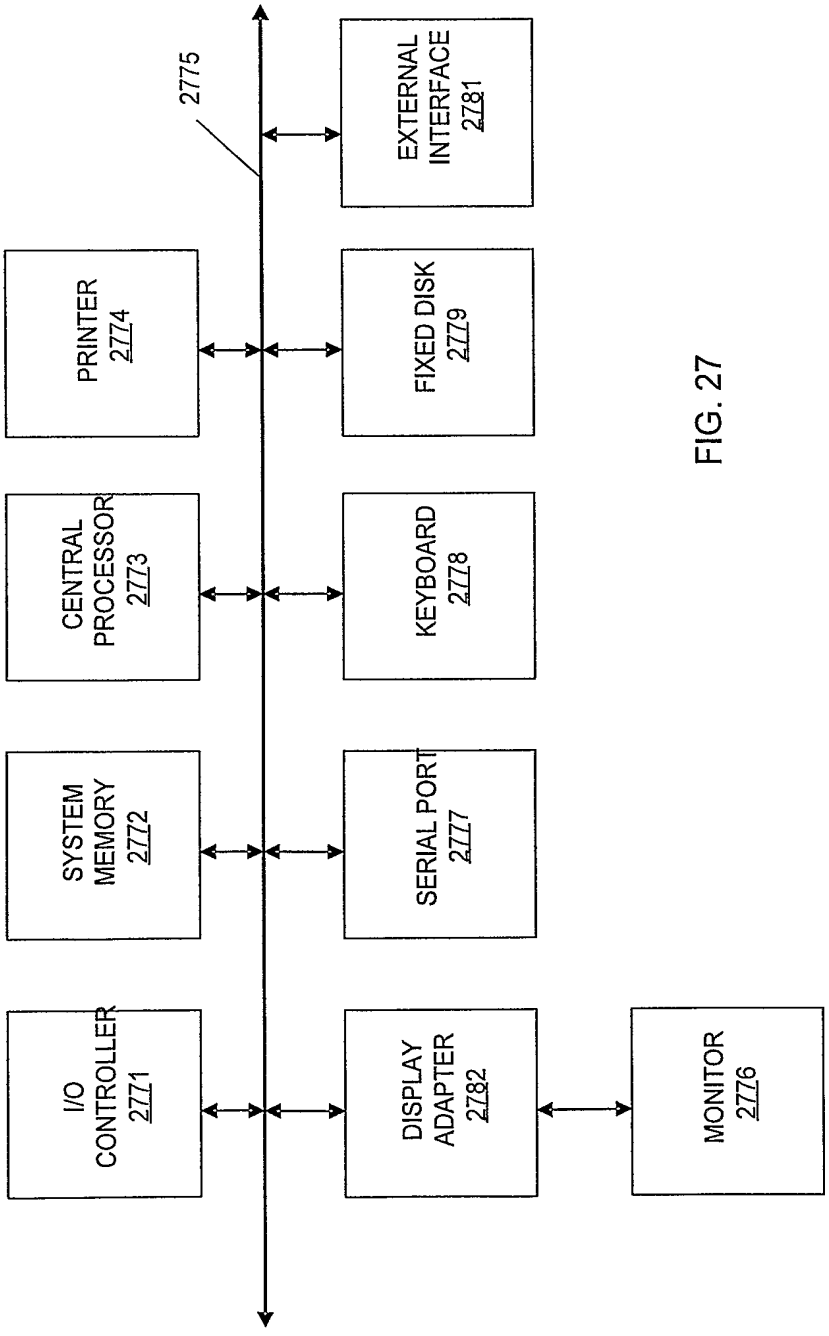

2775
EXTERNAL INTERFACE 2781
PRINTER 2774
FIXED DISK 2779
CENTRAL PROCESSOR 2773
KEYBOARD 2778
SYSTEM MEMORY 2772
SERIAL PORT 2777
I/O CONTROLLER 2771
DISPLAY ADAPTER 2782
MONITOR 2776

FIG. 27

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2008/002524

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/00 C12Q1/68

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)
G06F 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 practical, search terms used)
EPO-Internal, BIOSIS, EMBASE, COMPENDEX, INSPEC, 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	ZHOU W ET AL: "Counting alleles to predict recurrence of early-stage colorectal cancers" LANCET THE, LANCET LIMITED. LONDON, GB, vol. 359, no. 9302, 19 January 2002 (2002-01-19), pages 219-225, XP004791874 ISSN: 0140-6736 cited in the application abstract page 220, right-hand column, line 38 - page 221, right-hand column, line 10 figures 1,2 ----- -/--	1-12, 14-18, 25,26, 30,31

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier document 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
24 October 2008	30/10/2008

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	Kürten, Ivayla

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

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SHIH I-M ET AL: "Evidence that genetic instability occurs at an early stage of colorectal tumorigenesis" CANCER RESEARCH, vol. 61, February 2002 (2002-02), pages 818-822, XP007905911 page 819, right-hand column, line 13 - page 820, right-hand column, line 21 figure 2	1-12, 14-18, 25,26, 30,31
X	TONG YU K ET AL: "Noninvasive prenatal detection of fetal trisomy 18 by epigenetic allelic ratio analysis in maternal plasma: Theoretical and empirical considerations" CLINICAL CHEMISTRY, AMERICAN ASSOCIATION FOR CLINICAL CHEMISTRY, WASHINGTON, DC, vol. 52, no. 12, 13 October 2006 (2006-10-13), pages 2194-2202, XP002470084 ISSN: 0009-9147 cited in the application	1-12, 14-18,31
A	page 2197, left-hand column, line 14 - page 2199, right-hand column, line 32 page 2201 figure 2	25,26,30
P,X	LO YMD ET AL: "Digital PCR for the molecular detection of fetal chromosomal aneuploidy." PNAS, vol. 104, no. 32, 7 August 2007 (2007-08-07), pages 13116-13121, XP007905909 page 13117, left-hand column - page 13118, right-hand column	1-12, 14-18, 25,26, 30,31
A	LO YMD ET AL: "Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection" NATURE MEDICINE, vol. 13, no. 2, February 2007 (2007-02), pages 218-223, XP007905910 cited in the application page 219, left-hand column, line 5 - line 25 page 219, right-hand column, line 26 - page 221, left-hand column page 222, right-hand column, line 13 - line 24	1-12, 14-18, 25,26, 30,31

-/--

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EL KAROUI N ET AL: "Getting more from digital SNP data" STATISTICS IN MEDICINE, vol. 25, 2006, pages 3124-3133, XP007905913 cited in the application page 3125, line 25 - page 3128, line 19	1-12, 14-18, 25,26, 30,31
A	LO YMD AND CHIU RWK: "Prenatal diagnosis: progress through plasma nucleic acids." NATURE REVIEWS GENETICS, vol. 8, January 2007 (2007-01), pages 71-77, XP007905874 cited in the application the whole document	1-12, 14-18, 25,26, 30,31

INTERNATIONAL SEARCH REPORT

International application No.
PCT/GB2008/002524

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: 13,19-24,27-29
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
see FURTHER INFORMATION sheet PCT/ISA/210

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers allsearchable claims.

2. ☒ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search reportcovers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box II.2

Claims Nos.: 13,19-24,27-29

No opinion is given for claims 13, 19 to 24 and 27 to 29 because the subject-matter of said claims is not clearly defined contrary to the requirements of Article 6 PCT. The reasons are as follows:

According to claim 13 "the cutoff value is based on a measurement of one of the first portions or a measurement of one of the second portions". This vague expression does not enable the skilled reader to determine how the cutoff value is calculated nor which quantities are used in the calculation of said cutoff value. The subject-matter of claim 13 is unclear to such an extent that no meaningful search could be carried out.

According to claim 19, the "calculating of the first cutoff value includes at least one of: sequential probability ratio testing, false discovery rates, confidence intervals, and receiver operating characteristic curves." These listed methods allow for hypothesis testing (i.e. sequence imbalance vs sequence balance) based on given cutoff values. It is unclear how these methods are used for calculating the cutoff value itself.

Claims 20 to 24 define a proportion P1 and a set of probability values and state that the method includes "calculating the first cutoff value from the first proportion P1" (claim 20). This information, however, is insufficient to define said cutoff value. Moreover, the calculation of P1 itself is unclear: according to claim 21, P1 is calculated "using the first probability and the second probability". It is unclear how exactly these probabilities are used in the calculation of P1.

In claim 22, it is unclear what is being addressed by "an expected ratio relative to the non-reference nucleic acid sequence". Consequently, the calculation of the "first probability" remains unclear as well.

Claim 24 defines yet another probability, "third probability" but does not specify the purpose of this probability. It is unclear why this probability is defined and how it is used in the determination of the cutoff value. Also, according to claim 24 "calculating the second probability ... includes assuming that first probability and second probability are independent". It is unclear how the independence of the first and the second probabilities is incorporated in the calculation of the second probability.

The wording of claims 19 to 24 does not enable the skilled person to determine how the cutoff value is calculated. Consequently, no search regarding the definition of the cutoff value could be carried out. Although the description is more specific regarding the calculation of the cutoff values, the relevant passages are spread across pages 22 to 26 and could not be interpreted unambiguously.

Claim 27 also attempts to define the first cutoff value by "determining

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

a first average concentration ...; and multiplying the first average concentration by a factor derived from the first percentage to obtain a second average concentration ...". The claim only specifies that the factor is derived from the first percentage, but does not specify how exactly said factor is calculated. Furthermore, the result of the claim is a second average concentration. The claim does not specify how this second average concentration is used in the calculation of the cutoff value.

Claims 28 and 29 further specify the average concentration but do not enlighten the procedure of calculating the cutoff value.

Consequently, the method of calculating said cutoff value is unclear to such an extent that could not be meaningfully searched.

The applicant's attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination before the EPO (see EPO Guideline C-VI, 8.2), should the problems which led to the Article 17(2)PCT declaration be overcome.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-12,14-18,31

A method for determining whether a nucleic acid sequence imbalance in a sample exists whereby the average concentration of a reference nucleic acid in a reaction is taken into account.

---

2. claims: 25-26,30

A method for determining whether a nucleic acid sequence imbalance in a sample exists whereby the sample consists of two or more types of cells and the percentage of the nucleic acid sequence from the first type of cells is taken into account.

I hereby certify that this correspondence is being filed with the United States Patent and Trademark Office via the Office electronic filing system on 02/02/2023.

KILPATRICK TOWNSEND & STOCKTON LLP

By: /Stephen Quast/

PATENT
Attorney Docket No.: 080015-1343271-028430US
Client Ref. No.: 19/MED/901 US Con2

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of: Yuk-Ming Dennis Lo et al. Application No.: 17/891,899 Filed: August 19, 2022 For: DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS Customer No.: 20350	Confirmation No.: 1331 Examiner: Not Yet Assigned Technology Center/Art Unit: Not Yet Assigned INFORMATION DISCLOSURE STATEMENT

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Commissioner:

The references cited on attached form PTO/SB/08A are being called to the attention of the Examiner. In accordance with the provisions of 37 CFR § 1.98(a)(2), copies of any cited U.S. Patents and U.S. Patent Application Publications are not provided. Copies of the remaining cited references are provided.

It is respectfully requested that the cited references be expressly considered during the prosecution of this application and the references be made of record therein and appear among the "references cited" on any patent to issue therefrom.

As provided for by 37 CFR § 1.97(g) and (h), no inference should be made that the information and references cited are prior art merely because they are in this statement and no

representation is being made that a search has been conducted or that this statement encompasses all the possible relevant information. This Information Disclosure Statement is being filed before the mailing date of the first Office Action on the merits.

Applicant believes that no fee is required for submission of this statement. However, if any fees are due for the submission of this Information Disclosure Statement, please deduct those fees from Deposit Account No. 20-1430.

Respectfully submitted,

/David D. Hsu/

David D. Hsu
Registration No. 71659

KILPATRICK TOWNSEND & STOCKTON LLP

Espacenet

Bibliographic data: KR101896167 (B1) — 2018-09-07

DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE

Inventor(s): 로, 욱-밍 데니스, ; 치우, 로사 와이 쿤, ; 찬, 콴 치, ; 지, 베니 충 잉, ; 총, 카 춘

Applicant(s): 더 차이니즈 유니버시티 오브 홍콩

Classification: - international: ***C12Q1/68; G06F19/18; G06F19/22***
- cooperative: **C12Q1/68 (EA)**; **C12Q1/6809 (CN)**; **C12Q1/6827 (EP, CN, KR, US)**; **C12Q1/6869 (CN)**; **C12Q1/6883 (EP, CN, KR, US)**; **C12Q1/6888 (KR, US)**; **G01N33/48 (EA)**; G01N33/50 (EA); **G16B20/00 (CN, KR, US)**; **G16B20/10 (EP)**; **G16B30/00 (CN, KR)**; C12Q2600/112 (KR, US); C12Q2600/154 (CN, KR, US); C12Q2600/156 (KR, US); G01N2800/387 (KR, US); G16B30/00 (EP, US); Y02A90/10 (EP, US) more

Application number: KR20167021213 20080723 Global Dossier

Priority number(s): US20070951438P 20070723 ; WO2008GB02524 20080723

Also published as: AU2008278839 (A1) AU2008278839 (B2) AU2008278843 (A1) BRPI0814670 (A2) BRPI0814670 (B1) CA2693081 (A1) CA2693081 (C) CA2694007 (A1) CA2694007 (C) CA2900927 (A1) CA2900927 (C) CA3009992 (A1) CA3029497 (A1) CA3076142 (A1) CA3076159 (A1) CN101849236 (A) CN101971178 (A) CN101971178 (B) CN103849684 (A) CN103853916 (A) CN103853916 (B) CN103902809 (A) CN103902809 (B) CN106676188 (A) CN106834474 (A) CN106834474 (B) CN106834481 (A) CN106886688 (A) CN106886688 (B) CN107083424 (A) CN107083425 (A) CY1114773 (T1) CY1117525 (T1) DK2183693 (T3) DK2183693 (T4) DK2183693 (T5) DK2514842 (T3) DK2527471 (T3) DK2557519 (T3) EA017966 (B1) EA028642 (B1) EA035451 (B1) EA035451 (B9) EA201000231 (A1) EA201201551 (A1) EA201300072 (A1) EA201600280 (A1) EA201791612 (A2) EA201791612 (A3) EP2183692 (A1) EP2183692 (B1) EP2183693 (A1) EP2183693 (B1) EP2183693 (B2) EP2514842 (A2) EP2514842 (A3) EP2514842 (B1) EP2527471 (A2) EP2527471 (A3) EP2527471 (B1) EP2557517 (A2) EP2557517 (A3) EP2557518 (A2) EP2557518 (A3) EP2557518 (B1) EP2557519 (A2) EP2557519 (A3)

EP2557519 (B1) EP2557520 (A2) EP2557520 (A3)
EP3067807 (A1) EP3540739 (A1) EP3656870 (A1)
EP3745405 (A1) EP3770275 (A1) ES2441807 (T3)
ES2441807 (T5) ES2571738 (T3) ES2792802 (T3)
HK1144024 (A1) HK1177768 (A1) HK1182195 (A1)
HK1199067 (A1) HK1224033 (A1) HRP20140009 (T1)
HRP20140009 (T4) HRP20160493 (T1) HUE030510 (T2)
IL203311 (A) IL233261 (A) JP2010534068 (A) JP2010534069 (A)
JP2014073134 (A) JP2015142588 (A) JP2016185162 (A)
JP2017148073 (A) JP2019000113 (A) JP2019013245 (A)
JP2020031663 (A) JP2020031664 (A) JP2020115887 (A)
JP5519500 (B2) JP5736170 (B2) JP6151739 (B2)
JP6383837 (B2) JP6522554 (B2) JP6629940 (B2)
JP6695392 (B2) KR101646978 (B1) KR101829564 (B1)
KR101829565 (B1) KR101916456 (B1) KR101966262 (B1)
KR101972994 (B1) KR102060911 (B1) KR102076438 (B1)
KR102112438 (B1) KR102128960 (B1) KR102147626 (B1)
KR102197512 (B1) KR20100058503 (A) KR20100075826 (A)
KR20160030404 (A) KR20160113145 (A) KR20160113146 (A)
KR20160113147 (A) KR20160113148 (A) KR20170127073 (A)
KR20180100713 (A) KR20180114251 (A) KR20180121695 (A)
KR20190114039 (A) KR20190114041 (A) KR20190143494 (A)
KR20200055151 (A) KR20200100860 (A) KR20210006468 (A)
MX2010000846 (A) MX341573 (B) MX346069 (B) NZ582702 (A)
NZ600407 (A) PL2183693 (T3) PL2183693 (T4) PL2183693 (T5)
PL2514842 (T3) PT2183693 (E) SG183062 (A1) SI2183693 (T1)
SI2183693 (T2) SI2514842 (T1) US10208348 (B2)
US2009029377 (A1) US2009087847 (A1) US2014045181 (A1)
US2014256559 (A1) US2014256560 (A1) US2014329695 (A1)
US2014329696 (A1) US2019136323 (A1) US2020056242 (A1)
US8706422 (B2) US8972202 (B2) US9051616 (B2)
WO2009013492 (A1) WO2009013496 (A1) ZA201000524 (B) less

Abstract not available for KR101896167 (B1)
Abstract of corresponding document: WO2009013492 (A1)

Methods, systems, and apparatus are provided for determining whether a nucleic acid sequence imbalance exists within a biological sample. One or more cutoff values for determining an imbalance of, for example, the ratio of the two sequences (or sets of sequences) are chosen. The cutoff value may be determined based at least in part on the percentage of fetal DNA in a sample, such as maternal plasma, containing a background of maternal nucleic acid sequences. The cutoff value may also be determined based on an average concentration of a sequence per reaction. In one aspect, the cutoff value is determined from a proportion of informative wells that are estimated to contain a particular nucleic acid sequence, where the proportion is determined based on the above-mentioned percentage and/or average concentration. The cutoff value may be determined using many different types of methods, such as sequential probability ratio testing (SPRT).

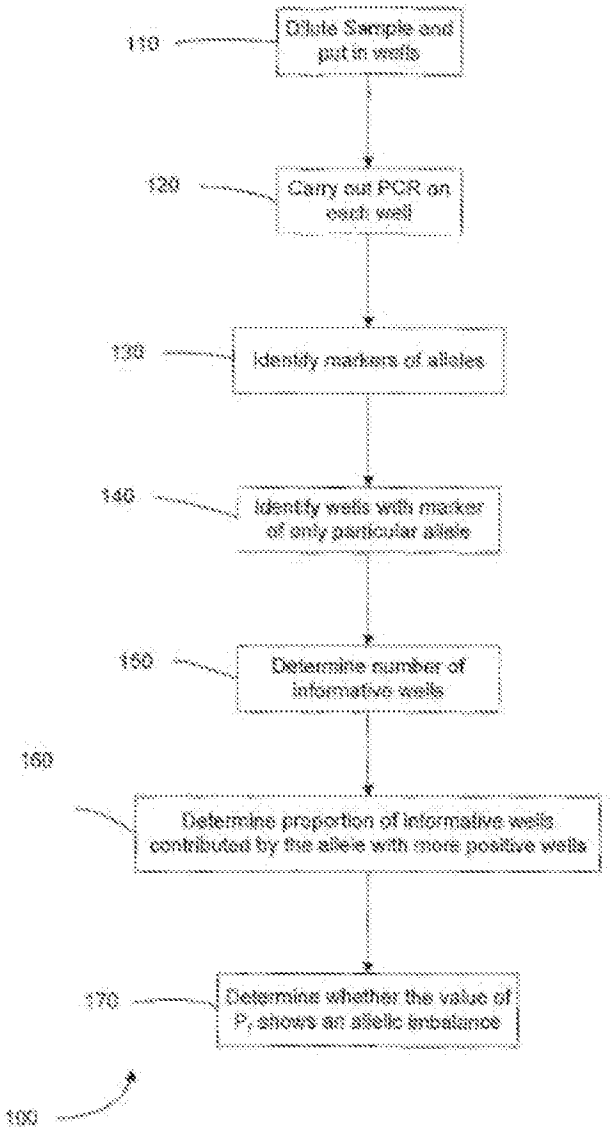

110
Dilute Sample and put in wells
120
Carry out PCR on each well
130
Identify markers of alleles
140
Identify wells with marker of only particular allele
150
Determine number of informative wells
160
Determine proportion of informative wells contributed by the allele with more positive wells
170
Determine whether the value of P_1 shows an allelic imbalance
100

FIG. 1

(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2018년09월07일
(11) 등록번호 10-1896167
(24) 등록일자 2018년08월31일

(51) 국제특허분류(Int. Cl.)
G06F 19/18 (2011.01) *C12Q 1/68* (2018.01)
G06F 19/22 (2017.01)
(52) CPC특허분류
G06F 19/18 (2013.01)
C12Q 1/6827 (2018.05)
(21) 출원번호 10-2016-7021213(분할)
(22) 출원일자(국제) 2008년07월23일
심사청구일자 2016년09월01일
(85) 번역문제출일자 2016년08월02일
(65) 공개번호 10-2016-0113147
(43) 공개일자 2016년09월28일
(62) 원출원 특허 10-2010-7003906
원출원일자(국제) 2008년07월23일
심사청구일자 2013년07월23일
(86) 국제출원번호 PCT/GB2008/002524
(87) 국제공개번호 WO 2009/013492
국제공개일자 2009년01월29일
(30) 우선권주장
60/951,438 2007년07월23일 미국(US)
(56) 선행기술조사문헌
JP2007515947 A*
KR1020020064298 A*
WO2007075826 A2*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
더 차이니즈 유니버시티 오브 홍콩
중국 홍콩 엔. 티. 세틴 더 차이니즈 유니버시티 오브 홍콩
(72) 발명자
로, 욕-밍 데니스
홍콩, 구룡, 호만틴, 7 킹 탁 스트리트, 4 플로어
치우, 로사 와이 큔
홍콩, 뉴 테리토리스, 타이 포, 1 홍 함 드라이브, 콘스텔레이션 코브, 블록 1, 플랫 1에이
(*공변에 계속*)
(74) 대리인
손민

전체 청구항 수 : 총 20 항 심사관 : 박권아

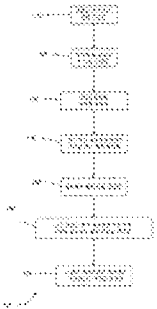
(54) 발명의 명칭 **핵산 서열 불균형의 결정**

(57) 요 약

본 발명에서는, 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하는 방법, 시스템 및 장치가 제공된다. 예를 들어, 2개의 서열 (또는 서열 세트)의 비의 불균형을 측정하기 위한 하나 이상의 분별값이 선택된다. 분별값은 적어도 부분적으로 샘플, 예컨대 모체 핵산 서열의 배경을 함유하는 모체 혈장 중 태아 DNA의 백

(*뒷면에 계속*)

대 표 도 - 도1



분율을 기준으로 결정될 수 있다. 분별값은 또한 반응 당 시열의 평균 농도를 기준으로 결정될 수 있다. 한 측면에서, 분별값은 특정 핵산 서열을 함유하는 것으로 평가되는 정보제공 핵의 비율로부터 결정되며, 그 비율은 상기 연결된 백분율 및/또는 평균 농도를 기준으로 결정된다. 분별값은 다수의 상이한 유형의 방법, 예컨대 순차 확률비 검증 (SPRT)을 사용하여 결정될 수 있다.

(52) CPC특허분류
C12Q 1/6883 (2018.05)
C12Q 1/6888 (2018.05)
G06F 19/22 (2018.05)
C12Q 2600/112 (2013.01)
C12Q 2600/154 (2013.01)
C12Q 2600/156 (2013.01)
G01N 2800/387 (2013.01)

(72) 발명자
찬, 환 치
홍콩, 구룡, 에이 푸 신 추엔, 프로스웨어 스트리트, 블록 34, 13/에프, 플랫 에이
지, 베니 총 잉
홍콩, 뉴 테리토리스, 마 온 산, 타 코스타, 타워 2, 플랫 18이
총, 카 춘
홍콩, 뉴 테리토리스, 타이 포, 푸 신 에스테이트, 신 경 하우스, 20/에프, 플랫 06

청 구 서

청구범위

청구항 1

핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 작업을 수행하기 위한 컴퓨터 시스템 제어용의 다수의 지시로 코딩된 비일시적 컴퓨터 판독가능 기록 매체로서, 상기 지시는 하기를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체:

다수의 반응으로부터 데이터를 수득하는 단계로, 상기 데이터는

(1) 생물학적 샘플 중의 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이터; 및

(2) 임상적으로 관련된 핵산 서열과 상이한 생물학적 샘플 중의 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이터를 포함하고,

상기 임상적으로 관련된 핵산 서열과 배경 핵산 서열이 제1 유형의 세포 및 하나 이상의 제2 유형의 세포로부터 유래된 것인, 단계;

두 데이터 세트로부터 파라미터를 결정하는 단계;

생물학적 샘플에서 제1 유형의 세포로부터의 핵산 서열의 양을 측정하여 수득된 제1 백분율로부터 제1 분별값(cutoff value)을 유도하는 단계;

파라미터를 제1 분별값과 비교하는 단계; 및

비교를 기준으로, 생물학적 샘플 내에 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계.

청구항 2

제1항에 있어서, 서열 불균형이 종양에서 염색체의 획득 또는 소실과 관련되는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 3

제1항에 있어서, 제1 유형의 세포가 제1 유기체로부터의 것이고, 제2 유형의 세포가 제2 유기체로부터의 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 4

제3항에 있어서, 제1 유기체가 제2 유기체인 태아를 임신한 여성 대상체이고, 제1 백분율의 결정이 여성 대상체와 태아 사이의 다형성 차이의 정량화를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 5

제3항에 있어서, 제1 유기체가 제2 유기체인 태아를 임신한 여성 대상체이고, 태아가 남성이며, 생물학적 샘플 중의 태아 DNA의 분획 농도를 결정하는 단계가 Y 염색체 DNA 농도를 결정하는 단계를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 6

제3항에 있어서, 제1 유기체가 제2 유기체인 태아를 임신한 여성 대상체이고, 생물학적 샘플 중의 태아 DNA의 분획 농도를 결정하는 단계가 제1 유전자좌에서의 태아-특이적 메틸화 패턴을 나타내는 DNA 분자의 양을 제1 유전자좌에서의 DNA 분자의 총량과 비교하는 단계를 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 7

제6항에 있어서, 태아-유도된 DNA 분자가 과-메틸화되고, 모체-유도된 DNA 분자가 저-메틸화되는 것인, 비일시

적 컴퓨터 판독가능 기록 매체.

청구항 8

제1항에 있어서, 상기 지시가

반응 당, 과소발현된 임상적으로 관련된 핵산 서열 또는 배경 핵산 서열인, 참조 핵산 서열의 제1 평균 농도를 결정하는 단계; 및

제1 평균 농도에 제1 백분율로부터 유도된 인자를 곱하여 참조 핵산 서열이 아닌 핵산 서열의 제2 평균 농도를 얻는 단계를 추가로 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 9

제3항에 있어서, 상기 지시가 참조 핵산 서열에 대한 데이터로부터 유도된 값이 입력된 역 확률 분포를 사용하여 다수의 반응 각각에서 참조 핵산 서열의 평균 농도를 결정하는 단계를 더 포함하는 것인, 비일시적 컴퓨터 판독가능 기록 매체.

청구항 10

태아를 임신한 여성 대상체의 생물학적 샘플 내의 태아 DNA의 분획 농도를 결정하는 방법으로서, 생물학적 샘플은 여성 대상체 및 태아로부터 핵산 분자를 포함하고, 모체는 제1 대립유전자에 대한 제1 유전자좌에서 동형접합성이고 태아는 제1 대립유전자 및 제1 대립유전자와 상이한 제2 대립유전자에 대한 제1 유전자좌에서 이형접합성이며, 상기 방법은 하기를 포함하는 것인, 방법:

생물학적 샘플로부터의 핵산 분자를 포함하는 제1의 다수의 반응으로부터 제1 데이터를 수득하는 단계로, 상기 반응은 다수의 관심 폴리뉴클레오티드 서열의 존재 또는 부재를 나타내고, 상기 데이터는

(1) 제1 대립유전자에 대해 양성인 반응의 제1 수를 나타내는 제1 세트의 정량 데이터; 및

(2) 제2 대립유전자에 대해 양성인 반응의 제2 수를 나타내는 제2 세트의 정량 데이터를 포함하는, 단계; 및

제1 수를 제2 수와 비교하여 태아 DNA의 분획 농도를 결정하는 단계.

청구항 11

제10항에 있어서, 태아 DNA의 분획 농도가 제1 수 및 제2 수를 사용하여 백분율로서 결정되는 것인 방법.

청구항 12

제10항에 있어서, 하기 단계를 더 포함하는 것인 방법:

생물학적 샘플로부터의 핵산 분자를 포함하는 제2의 다수의 반응으로부터 제2 데이터를 수득하는 단계로, 상기 데이터는

(1) 제1 유전자좌를 포함하지 않는 하나 이상의 임상적으로 관련된 핵산 서열의 제3 양을 나타내는 제3 세트의 정량 데이터; 및

(2) 하나 이상의 임상적으로 관련된 핵산 서열과 상이한 하나 이상의 배경 핵산 서열의 제4 양을 나타내는 제4 세트의 정량 데이터를 포함하고,

상기 임상적으로 관련된 핵산 서열 및 배경 핵산 서열은 여성 대상체 및 태아로부터의 핵산 분자로부터 유래된 것인, 단계;

제3 및 제4 데이터 세트로부터 제3 양과 제4 양 사이의 상대적 양을 제공하는 파라미터를 결정하는 단계;

태아 DNA의 분획 농도로부터 제1 분별값을 유도하는 단계;

파라미터를 제1 분별값과 비교하는 단계; 및

비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계.

청구항 13

제12항에 있어서, 서열 불균형이 염색체 이수성(aneuploidy)인 방법.

청구항 14

제12항에 있어서, 임상적으로 관련된 핵산 서열 중 하나가 유전자 다형성의 대립유전자이고, 배경 핵산 서열 중 하나가 유전자 다형성의 또 다른 대립유전자인 방법.

청구항 15

제12항에 있어서, 제2의 다수의 반응을 수행하기 전에 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열에 대해 생물학적 샘플을 농축시키는 단계를 더 포함하는 방법.

청구항 16

제15항에 있어서, 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열에 대해 생물학적 샘플을 농축시키는 단계가 혼성화로 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열을 선별하기 위해 올리고뉴클레오타드 어레이를 사용하는 단계를 포함하는 것인 방법.

청구항 17

제15항에 있어서, 생물학적 샘플을 농축시키는 단계가 하나 이상의 임상적으로 관련된 핵산 서열 및 하나 이상의 배경 핵산 서열로부터 DNA를 증폭시키는 단계를 포함하는 것인 방법.

청구항 18

제12항에 있어서, 태아 DNA의 분획 농도가 파라미터 결정 후 결정되는 것인 방법.

청구항 19

제12항에 있어서, 제1의 다수의 반응이 제2의 다수의 반응과 동일한 것인 방법.

청구항 20

제10항에 있어서, 반응이 서열분석 반응 또는 중쪽 반응인 방법.

발명의 설명

기 술 분 야

[0001] 우선권 주장

[0002] 본 출원은 2007년 7월 23일 출원된, 발명의 명칭이 "핵산 서열 불균형의 결정(DETERMINING A NUCLEIC ACID SEQUENCE IMBALANCE)"인 미국 가출원 번호 제60/951438호 (사건관리 번호 제016285-006200US호)를 우선권으로 주장하여 이의 정규출원이고, 상기 가출원의 전체 내용은 모든 목적을 위해 참조로 본원에 포함된다.

[0003] 관련 출원의 상호 참조

[0004] 본 출원은 또한 동시에 출원된, 발명의 명칭이 "개눔 서열분석을 사용한 태아 염색체 이수성의 진단(DIAGNOSING FETAL CHROMOSOMAL ANEUPLOIDY USING GENOMIC SEQUENCING)"인 정규출원 (사건관리 번호 제016285-006220US호)에 관한 것이고, 이 가출원의 전체 내용은 모든 목적을 위해 참조로 본원에 포함된다.

[0005] 발명의 분야

[0006] 본 발명은 일반적으로 2개의 상이한 핵산 서열간 불균형을 결정하는 유전자형 및 질환의 진단 시험, 보다 특히 모체 혈액 샘플의 시험을 통한 태아에서의 다운 증후군, 다른 염색체 이수성(aneuploidy), 돌연변이 및 유전자형의 확인에 관한 것이다. 본 발명은 또한 암의 검출, 이식의 모니터링 및 감염성 질환의 모니터링에 관한 것이다.

해 명 기 술

[0008] 유전 질환, 암 및 다른 상태는 종종 2개의 상응하는 염색체 또는 대립유전자 또는 다른 핵산 서열에서의 불균형으로부터 초래되거나 이러한 불균형을 생성한다. 즉, 다른 서열에 비해 한 서열의 양이 정상적인 경우보다 많거나 적다. 통상적으로, 정상 비율은 대등한 50/50 비율이다. 다운 증후군 (삼염색체 21)은 추가 염색체 21의 불균형을 갖는 질환이다.

[0009] 삼염색체 21의 통상적인 출생 전 진단 방법은 한정된 태아 소실 위험을 갖는 침습적 절차, 예컨대 양수검사 또는 융모막 융모 생검에 의해 태아 물질을 샘플링하는 것을 포함한다. 비-침습적 절차, 예컨대 초음파검사 및 생화학 마커에 의한 스크리닝은 한정적인 침습적 진단 절차 전에 임신 여성을 위험-등급화하는데 사용되었다. 그러나, 이러한 스크리닝 방법은 전형적으로 핵심 염색체 이상 대신에 삼염색체 21과 관련된 부대증후를 측정하므로, 최적에 못미치는 진단 정확성 및 다른 단점들 (예컨대 재태 기간에 의해 많은 영향을 받음)을 갖는다.

[0010] 1997년에 모체 혈장 중 순환 무-세포 태아 DNA의 발견은 비-침습적 출생 전 진단의 새로운 가능성을 제공하였다 (문헌 [Lo, YMD and Chiu, RWK 2007 Nat Rev Genet 8, 71-77]). 이 방법이 성별-연계된 유전자 장애 (문헌 [Costa, JM et al. 2002 N Engl J Med 346, 1502]) 및 특정 단일 유전자 장애 (문헌 [Lo, YMD et al. 1998 N Engl J Med 339, 1734-1738])의 출생 전 진단에 용이하게 적용되었지만, 태아 염색체 이수성의 출생 전 검출에 적용하기에는 상당한 난점이 있었다 (상기 문헌 [Lo, YMD and Chiu, RWK 2007]). 첫째, 태아 핵산은 모체 혈장 중에서, 종종 분석에 방해가 될 수 있는 다량의 모체 기원 배경 핵산과 공존한다 (문헌 [Lo, YMD et al. 1998 Am J Hum Genet 62, 768-775]). 둘째, 태아 핵산은 모체 혈장 중에서 우세하게는 무-세포 형태로 순환하여, 태아 게놈 내 유전자 또는 염색체의 무어 정보 유도를 어렵게 만든다.

[0011] 이들 난점을 극복하는 유의한 개발이 최근에 이루어졌다 (문헌 [Benachi, A & Costa, JM 2007 Lancet 369, 440-442]). 한 접근법은 모체 혈장 중 태아-특이적 핵산을 검출하여, 모체 배경 방해 문제를 극복하는 것이다 (상기 문헌 [Lo, YMD and Chiu, RWK 2007]). 염색체 21의 무어는 태반-유래 DNA/RNA 분자에서 다형성 대립유전자의 비율로부터 추론되었다. 그러나, 이러한 방법은, 샘플이 보다 적은 양의 표적 유전자를 함유하는 경우 덜 정확하며, 하나의 다형성이 사용되는 경우 단지 그 집단의 아군인, 표적 다형성에 대해 이형접합성인 태아에게만 적용될 수 있다.

[0012] 알란 등(Dhallan et al)의 문헌 [Dhallan, R. et al. 2007, supra Dhallan, R. et al. 2007 Lancet 369, 474-481]에는, 모체 혈장에 포름알데히드를 첨가함으로써 순환 태아 DNA의 비율을 증가시키는 별법의 전략이 기재되어 있다. 모체 혈장 중 태아에 의해 기여된 염색체 21 서열의 비율은, 염색체 21 상의 단일 뉴클레오티드 다형성 (SNP)에 대한 무계-유전 태아-특이적 대립유전자 대 비-태아-특이적 대립유전자의 비율을 평가함으로써 결정되었다. SNP 비율은 참조 염색체에 대해 유사하게 추정되었다. 이어서, 태아 염색체 21의 불균형은, 염색체 21에 대한 SNP 비율과 참조 염색체에 대한 SNP 비율 사이의 통계적으로 유의한 차이를 검출함으로써 추론되었으며, 상기 유의성은 0.05 이하의 고정 p-값을 사용하여 정의된다. 고 집단 적용범위를 보장하기 위해, 염색체당 500개 초과의 SNP가 표적화되었다. 그러나, 고비율로 증가시키기 위한 포름알데히드의 효과와 관련하여 논쟁이 있었으므로 (문헌 [Chung, GTY, et al. 2005 Clin Chem 51, 655-658]), 방법의 재현성은 더 평가될 필요가 있다. 또한, 각각의 태아 및 모체는 각각의 염색체에 대해 상이한 수의 SNP의 정보를 제공하므로, SNP 비율 비교를 위한 통계적 시험의 영향력은 건 마다 변할 것이다 (문헌 [Lo, YMD & Chiu, RWK. 2007 Lancet 369, 1907]). 또한, 이들 접근법은 유전자 다형성의 검출에 좌우되므로, 이들 다형성에 이형접합성인 태아에게 국한된다.

[0013] 삼염색체 21 및 정배수체 태아로부터 수득한 양수세포 배양물에서 염색체 21 유전자좌 및 참조 유전자좌의 중합효소 연쇄 반응 (PCR) 및 DNA 정량화를 사용하여, 짐메르만 등(Zimmermann et al) (문헌 [2002 Clin Chem 48, 362-363])은 삼염색체 21 태아에서 염색체 21 DNA 서열의 1.5배 증가를 기준으로 상기 2군의 태아를 구별할 수 있었다. DNA 주형 농도의 2배 차이는 단지 한 임계 사이클(threshold cycle (Ct))의 차이를 구성하기 때문에, 1.5배 차이의 구별은 통상적인 실시간 PCR의 한계였다. 보다 미세한 정도의 정량 구별을 달성하기 위해, 별법의 전략이 필요하다. 따라서, 본 발명의 일부 실시양태는 이러한 목적을 위해 디지탈 PCR (문헌 [Vogelstein, B et al. 1999 Proc Natl Acad Sci USA 96, 9236-9241])을 사용한다.

[0014] 디지탈 PCR은 핵산 샘플에서 대립유전자 비율이 벗어난 것을 검출하기 위해 개발되었다 (문헌 [Chang, HW et al. 2002 J Natl Cancer Inst 94, 1697-1703]). 일상적으로, 디지탈 PCR은 종양 DNA 샘플에서 이형접합성의 소실 (LOH)을 검출하는데 유용한 것으로 나타났다 (문헌 [Zhou, W. et al. 2002 Lancet 359, 219-225]). 디지

알 PCR 결과의 분석을 위해, 순차 확률비 검증 (SPRT)이 이전 연구에 의해 제안되어, 샘플 중 LOH의 존재를 제시하는 것과 제시하지 않는 것으로 실험 결과를 분류하였다 (문헌 [El Karoui et al. 2006 Stat Med 25, 3124-3133]). 이전 연구에서 사용된 방법에서, LOH를 결정하기 위한 문별값 (cutoff value)은 DNA 중 2개 대립유전자의 고정 참조 비율 2/3을 사용하였다. 모체 혈장 중 태아 핵산의 양, 비율 및 농도는 변할 수 있기 때문에, 이들 방법은 모체 혈장 중 모체 핵산의 배경에서 태아 핵산을 사용하여 삼염색체 21을 검출하는데 적합하지 않다.

[0015] 순환 태아 핵산 분석을 기준으로 하는 태아 삼염색체 21 (및 다른 불균형) 검출을 위한 비-침습적 시험, 특히 유전자 다형성 및/또는 태아-특이적 마커의 사용에 독립적인 비-침습적 시험을 수행하는 것이 바람직하다. 또한, 정확한 문별값 결정 및 시험 제수를 수행하여, 정확성에 필요한 모체 혈장 핵산 분자의 양 및/또는 태아타의 횟 수를 감소시킬 수 있음으로써, 증가된 효율 및 비용-효과를 제공하는 것이 바람직하다. 또한, 비-침습적 시험이 고 감수성 및 특이성을 가져 오류 진단을 최소화하는 것이 바람직하다.

[0016] 모체 혈장 중 태아 DNA 검출에 대한 모다든 적용은 베타-지중해빈혈과 같은 단일 유전자 장애의 출생 전 진단에 대한 것이다. 그러나, 태아 DNA는 단지 모체 혈장 중 소수 분획의 DNA를 구성하므로, 이러한 접근법은 단지 태아에서 부계로부터 유전되었지만 모계로부터는 존재하지 않는 돌연변이를 검출할 수 있을 뿐이라고 생각된다. 이러한 예로는, 베타-지중해빈형을 일으키는 베타-글로빈 유전자의 코돈 41/42에서의 4 bp 결실 (문헌 [Chiu RWK et al. 2002 Lancet, 360, 998-1000]) 및 낭포성 섬유증을 일으키는 낭포성 섬유증 막횡단 전도 조절자 유전자의 Q890X 돌연변이 (문헌 [Gonzalez-Gonzalez et al 2002 Prenat Diagn, 22, 946-8])가 포함된다. 그러나, 베타-지중해빈형 및 낭포성 섬유증 둘 다는, 이 질환이 그 자체로 표출되기 전에 태아에게 각각의 부모로부터 돌연변이가 유전될 필요가 있는 상염색체 열성 상태이므로, 부계-유전 돌연변이만의 검출은 단지 상기 질환에 걸린 태아를 가질 위험을 25%에서 50%로 증가시킬 것이다. 진단적으로 이는 이상적이 아니다. 따라서, 현존하는 접근법의 주된 진단 적용은, 부계-유전된 태아 돌연변이가 모체 혈장에서 검출될 수 없는 경우, 이후 태아가 동형접합성 질환 상태에 걸리는 것으로부터 배제될 수 있는 경우의 시나리오를 위해서일 것이다. 그러나, 진단적으로, 이러한 접근법은 부계 돌연변이의 음성 검출을 기준으로 결론을 내린다는 단점을 갖는다. 따라서, 상기 제한 없이, 완전한 태아 유전자형 (동형접합성 정상, 동형접합성 돌연변이 또는 이형접합성)이 모체 혈장으로부터 결정되도록 하는 접근법이 매우 바람직할 것이다.

발명의 내용

해결하려는 과제

[0018] 개요.

[0019] 본 발명의 실시양태는, 핵산 서열 불균형 (예를 들어, 대립유전자 불균형, 돌연변이 불균형 또는 염색체 불균형)이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 방법, 시스템 및 장치를 제공한다. 예를 들어, 2개의 서열 (또는 서열 세트) 양의 비율의 불균형을 결정하기 위한 하나 이상의 문별값이 선택된다.

[0020] 한 실시양태에서, 문별값은 적어도 부분적으로 생물학적 샘플, 예컨대 모체 핵산 서열의 배경을 함유하는 모체 혈장 또는 혈청 또는 뇨 중 태아 (임상적으로 관련된 핵산) 서열의 백분율을 기준으로 결정된다. 또다른 실시양태에서, 문별값은 다수의 반응에서 서열의 평균 농도를 기준으로 결정된다. 한 측면에서, 문별값은 특정 핵산 서열을 함유하는 것으로 평가되는 집단제공 핵의 비율로부터 결정되며, 그 비율은 상기 언급된 백분율 및/또는 평균 농도를 기준으로 결정된다.

[0021] 문별값은 다수의 상이한 유형의 방법, 예컨대 SPRT, 오류 발견, 신뢰 구간, 수신자 판단 특성 (receiver operating characteristic, ROC)을 사용하여 결정될 수 있다. 이러한 전략은 신뢰성 있는 분류가 만들어질 수 있기 전에 필요한 시험의 양을 더 최소화시켰다. 이는 주형 양이 종종 제한적인 혈장 핵산 분석과 특히 관련되어 있다.

[0022] 한 예시적 실시양태에 따라, (1) 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이타; 및 (2) 임상적으로 관련된 핵산 서열과 상이한 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이타를 포함하는 데이타를 다수의 반응으로부터 수득하는 단계; 두 데이타 세트로부터 파라미터를 결정하는 단계; 다수의 반응 각각에서, 임상적으로 관련된 핵산 서열 또는 배경 핵산 서열인 참조 핵산 서열의 평균 농도로부터 제1 문별값을 유도하는 단계; 파라미터를 제1 문별값과 비교하는 단계; 및 비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계를 포함하는, 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를

결정하기 위한 방법이 제공된다.

[0023] 또다른 예시적 실시양태에 따라, (1) 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이타; 및 (2) 임상적으로 관련된 핵산 서열과 상이한 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이타를 포함하며, 임상적으로 관련된 핵산 서열 및 배경 핵산 서열은 제1 유형 세포 및 하나 이상의 제2 유형 세포로부터 유래된 것인 데이타를 다수의 반응으로부터 수득하는 단계; 두 데이타 세트로부터 파라미터를 결정하는 단계; 생물학적 샘플에서, 제1 유형 세포로부터의 핵산 서열의 양을 측정하여 생성된 제1 백분율로부터 제1 분별값을 유도하는 단계; 파라미터를 분별값과 비교하는 단계; 및 비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계를 포함하는, 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 방법이 제공된다.

[0024] 본 발명의 다른 실시양태는 본원에 기재된 방법과 관련된 시스템 및 컴퓨터 판독가능 매체에 관한 것이다. 예시적 실시양태에 따르면, 본 발명은 핵산 서열 불균형이 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 작업을 수행하기 위한 컴퓨터 시스템 제어용의 다수의 지시로 코딩된, 비일시적 컴퓨터 판독가능 기록 매체에 관한 것으로서, 상기 지시는 (1) 생물학적 샘플 중의 임상적으로 관련된 핵산 서열의 제1 양을 나타내는 제1 세트의 정량 데이타; 및 (2) 임상적으로 관련된 핵산 서열과 상이한 생물학적 샘플 중의 배경 핵산 서열의 제2 양을 나타내는 제2 세트의 정량 데이타를 포함하고, 상기 임상적으로 관련된 핵산 서열과 배경 핵산 서열이 제1 유형의 세포 및 하나 이상의 제2 유형의 세포로부터 유래된 것인, 다수의 반응으로부터 데이타를 수득하는 단계; 두 데이타 세트로부터 파라미터를 결정하는 단계; 생물학적 샘플에서 제1 유형의 세포로부터의 핵산 서열의 양을 측정하여 수득된 제1 백분율로부터 제1 분별값을 유도하는 단계; 파라미터를 제1 분별값과 비교하는 단계; 및 비교를 기준으로, 핵산 서열 불균형이 존재하는지 여부의 분류를 결정하는 단계를 포함한다.

[0025] 하기 상세한 설명 및 수반되는 도면을 참조로 하여 본 발명의 성질 및 이점을 보다 잘 이해할 수 있다.

과제의 해결 수단

[0027] 정의

[0028] 본원에 사용된 용어 "생물학적 샘플"은 대상체 (예를 들어, 임신 여성과 같은 인간)로부터 취하여 하나 이상의 관심 핵산 분자(들)을 함유하는 임의의 샘플을 지칭한다.

[0029] 용어 "핵산" 또는 "폴리뉴클레오티드" 단일가닥 또는 이중가닥 형태의 데옥시리보핵산 (DNA) 또는 리보핵산 (RNA) 및 이들의 중합체를 지칭한다. 구체적으로 한정하지 않는 한, 상기 용어는 참조 핵산과 유사한 결합 성질을 가지며 천연 발생 뉴클레오티드와 유사한 방식으로 대사되는 천연 뉴클레오티드의 공지된 유사체를 함유하는 핵산을 포함한다. 달리 나타내지 않는 한, 특정 핵산 서열은 또한 암묵적으로, 그의 보존적으로 변형된 변이체 (예를 들어, 퇴행 코돈 치환체), 대립유전자, 오소로그(ortholog), SNP 및 상보적 서열, 및 명백하게 나타내어진 서열을 포함한다. 구체적으로, 퇴행 코돈 치환체는 하나 이상의 선택된 (또는 모든) 코돈의 제3 위치가 혼합 염기 및/또는 데옥시이노신 잔기로 치환되는 서열을 생성함으로써 달성될 수 있다 (문헌 [Batzer et al., Nucleic Acid Res. 19:5081 (1991)], [Ohtsuka et al., J. Biol. Chem. 260:2605-2608 (1985)] 및 [Rossolini et al., Mol. Cell. Probes 8:91-98 (1994)]). 용어 핵산은 유전자, cDNA, mRNA, 소형 비코딩 RNA, 마이크로 RNA (miRNA), 피위(Piwi)-상호작용 RNA, 및 유전자 또는 유전자좌에 의해 코딩되는 단형 해어핀 RNA (shRNA)와 상호교환적으로 사용된다.

[0030] 용어 "유전자"는 폴리펩티드 쇄를 생성하는데 관여하는 DNA 절편을 의미한다. 코딩 영역 전 및 후 영역 (리더 및 트레일러) 뿐만 아니라 개별 코딩 절편 (엑손) 사이의 삽입 서열 (인트론)을 포함할 수 있다.

[0031] 본원에 사용된 용어 "반응"은 관심있는 특정 폴리뉴클레오티드 서열의 존재 또는 부재를 나타내는 화학적, 효소적 또는 물리적 작용에 관여하는 임의의 과정을 지칭한다. "반응"의 한 예는 중합효소 연쇄 반응 (PCR)과 같은 증폭 반응이다. "반응"의 또다른 예는 합성 또는 라이게이션에 의한 서열분석 반응이다. "경쟁제공 반응"은 관심있는 하나 이상의 특정 폴리뉴클레오티드 서열의 존재를 나타내며 단지 하나의 관심 서열이 존재하는 환가지 경우에서의 반응이다. 본원에 사용된 용어 "웰"은 제한된 구조, 예를 들어 웰-형상 바이알, 웰, 또는 PCR 어레이 웰 내의 예정된 위치에서의 반응을 지칭한다.

[0032] 본원에 사용된 용어 "임상적으로 관련된 핵산 서열"은 통계적 불균형에 대해 시험될 보다 큰 계놈 서열의 절편에 상응하는 폴리뉴클레오티드 서열 또는 보다 큰 계놈 서열 그 자체를 지칭할 수 있다. 한 예는 염색체 21의 서열이다. 다른 예는 염색체 18, 13, X 및 Y를 포함한다. 또다른 예는 돌연변이 유전자 서열 또는 유전자 다

정성 또는 부모의 한쪽 또는 양쪽으로부터 태아에게 유전될 수 있는 카피 수 변형을 포함한다. 또다른 예는 악성 종양에서 돌연변이, 결실 또는 증폭된 서열, 예를 들어 이형접합성 소실 또는 유전자 복제가 일어나는 서열을 포함한다. 일부 실시양태에서, 다수의 임상적으로 관련된 핵산 서열 또는 동등하게는 임상적으로 관련된 핵산 서열의 다수의 마커는 불균형 검출을 위한 데이타를 제공하는데 사용될 수 있다. 예를 들어, 염색체 21 상의 5개의 비-연속 서열로부터의 데이타는 가능한 염색체 21 불균형의 결정을 위해 추가적인 방식으로 사용되어, 필요한 샘플 부피를 1/5로 유효하게 감소시킬 수 있다.

[0033] 본원에 사용된 용어 "배경 핵산 서열"은 임상적으로 관련된 핵산 서열에 대한 정상 비율이 공지되어 있는 (예를 들어 1 대 1 비율인) 핵산 서열을 지칭한다. 한 예로서, 배경 핵산 서열 및 임상적으로 관련된 핵산 서열은 이형접합성으로 인해 구별되는, 동일한 염색체로부터의 2개의 대립유전자이다. 또다른 예로, 배경 핵산 서열은 임상적으로 관련된 핵산 서열인 다른 대립유전자에 대해 이형접합성인 한 대립유전자이다. 또한, 각각의 배경 핵산 서열 및 임상적으로 관련된 핵산 서열 중 일부는 상이한 개체로부터 유래될 수 있다.

[0034] 본원에 사용된 용어 "참조 핵산 서열"은 반응 당 평균 농도가 공지되어 있거나 동등하게는 측정되는 핵산 서열을 지칭한다.

[0035] 본원에 사용된 용어 "과다표출된(overrepresented) 핵산 서열"은 2개의 핵산 서열 (예를 들어, 임상적으로 관련된 서열 및 배경 서열) 중 생물학적 샘플에서 다른 서열보다 풍부한 핵산 서열을 지칭한다.

[0036] 본원에 사용된 용어 "~기준으로"는 "적어도 부분적으로 ~기준으로"를 의미하며, 방법의 투입과 방법의 산출의 관계에서 일어나는, 한 값 (또는 결과)이 또다른 값의 결정에 사용되는 것을 지칭한다. 본원에 사용된 용어 "유도"는 그 유도가 식의 계산인 경우에 일어나는, 방법의 투입과 방법의 산출의 관계를 지칭한다.

[0037] 본원에 사용된 용어 "정량 데이타"는 하나 이상의 수치 값을 제공하는 하나 이상의 반응으로부터 얻은 데이타를 의미한다. 예를 들어, 특정 서열에 대한 형광 마커를 나타내는 웰의 수는 정량 데이타일 것이다.

[0038] 본원에 사용된 용어 "파라미터"는 정량 데이타 세트, 및/또는 정량 데이타 세트들 사이의 수치 관계를 특성화하는 수치 값을 의미한다. 예를 들어, 제1 핵산 서열의 제1 양과 제2 핵산 서열의 제2 양 사이의 비율 (또는 비율의 함수)이 파라미터이다.

[0039] 본원에 사용된 용어 "분별값"은 생물학적 샘플에 대한 분류의 2가지 이상의 상태 (예를 들어 질환 상태 및 비질환 상태)를 구별하는데 사용되는 수치 값을 의미한다. 예를 들어, 파라미터가 분별값보다 큰 경우 정량 데이타의 제1 분류가 만들어지거나 (예를 들어 질환 상태), 또는 파라미터가 분별값보다 작은 경우 정량 데이타의 상이한 분류가 만들어진다 (예를 들어 비질환 상태).

[0040] 본원에 사용된 용어 "불균형"은 하나 이상의 분별값으로 정의되는, 참조 정량으로부터 임상적으로 관련된 핵산 서열의 정량의 임의의 유의한 편차를 의미한다. 예를 들어, 참조 정량이 3/5의 비율일 수 있고, 따라서 측정 비율이 1:1인 경우에 불균형이 일어날 것이다.

[0042] 발명의 상세한 설명

[0043] 본 발명은, 다른 비-임상적으로 관련된 서열 (예를 들어, 염색체 또는 대립유전자 불균형)에 관하여 임상적으로 관련된 핵산 서열의 참조 (예를 들어 비질환 상태) 양에 비교되는 증가 또는 감소가 생물학적 샘플 내에 존재하는지 여부를 결정하기 위한 방법, 시스템 및 장치를 제공한다. 예를 들어 2개의 서열 (또는 서열 세트) 양의 비율과 관련하여 참조 양에 비교되는 변화 (즉 불균형)이 존재하는지 여부를 결정하기 위해 하나 이상의 분별값이 선택된다. 참조 양에서 검출된 변화는 다른 비-임상적 관련 서열에 대한 임상적 관련 핵산 서열에 관하여 임의의 편차 (상향 또는 하향)일 수 있다. 따라서, 참조 상태는 임의의 비율 또는 다른 양 (예를 들어 1-1 대응 이외의 것)일 수 있고, 변화를 의미하는 측정 상태는 하나 이상의 분별값에 의해 결정된 바와 같이 참조 양과 상이한 다른 양 또는 임의의 비율일 수 있다.

[0044] 임상적으로 관련된 핵산 서열 및 배경 핵산 서열은 제1 유형 세포 및 하나 이상의 제2 유형 세포로부터 유래될 수 있다. 예를 들어, 태아/태반 세포로부터 유래된 태아 핵산 서열은 생물학적 샘플, 예컨대 모체 세포로부터 유래된 모체 핵산 서열의 배경을 함유하는 모체 혈장 중에 존재한다. 따라서, 한 실시양태에서, 분별값은 적어도 부분적으로, 생물학적 샘플 중 제1 유형 세포의 백분율을 기준으로 결정된다. 샘플 중 태아 서열의 백분율은 임의의 태아-유래 유전자좌에 의해 결정될 수 있으며, 임상적으로 관련된 핵산 서열을 측정하는 것에 국한되지 않는다는 것을 주목한다. 또다른 실시양태에서, 분별값은 적어도 부분적으로, 생물학적 샘플, 예컨대 제1 비-악성 세포로부터 유래된 핵산 서열의 배경을 함유하는 혈장, 혈청, 타액 또는 뇨 중 종양 서열의 백분율을

기준으로 결정된다.

[0045] 또다른 실시양태에서, 분별값은 다수의 반응에서 서열의 평균 농도를 기준으로 결정된다. 한 측면에서, 분별값은 특정 핵산 서열을 함유하는 것으로 평가되는 정보내용 풀의 비율로부터 결정되며, 그 비율은 상기 언급된 배분율 및/또는 평균 농도를 기준으로 결정된다. 분별값은 다수의 상이한 유형의 방법, 예컨대 SPRT, 오류 방견, 신뢰 구간, 수신자 판단 특성 (ROC)을 사용하여 결정될 수 있다. 이러한 전략은 신뢰성 있는 분류가 만들어질 수 있기 전에 필요한 서열의 양을 더 최소화시킨다. 이는 주형 양이 종종 제한적인 혈장 핵산 분석과 특히 관련되어 있다. 디지털 PCR이 제시되더라도, 다른 방법이 사용될 수 있다.

[0046] 디지탈 PCR은 대부분의 양성 종목이 단일 주형 분자로부터의 신호를 반영하도록 매우 희석된 핵산에 대한 멀티플렉스 PCR 분석을 포함한다. 이로써 디지탈 PCR은 개별 주형 분자의 계수를 가능하게 한다. 분석된 PCR의 총 수 중 양성 종목의 비율로 인해 또는 비-희석 샘플 중의 주형 농도를 평가할 수 있다. 이러한 기술은 다양한 유전 현상의 검출을 가능하게 하도록 제시되었으며 (상기 문헌 [Vogelstein, B et al. 1999]), 종양 샘플 중 이형접합성 소실의 검출 (상기 문헌 [Zhou, W. et al. 2002]) 및 암 환자의 혈장 중 이형접합성 소실의 검출 (상기 문헌 [Chang, HW et al. 2002])을 위해 이전부터 사용되었다. 디지탈 PCR에 의한 주형 분자 정량화는 리포터 염료와 핵산 농도 사이의 루어랑-반응 관계에 의존하지 않기 때문에, 그의 분석 정확도는 실시간 PCR의 분석 정확도보다 이론적으로 월등하여야 한다. 따라서, 디지탈 PCR은 표적 유전자좌와 참조 유전자좌 사이의 보다 미세한 정도의 정량 차이를 검색적으로 구별할 수 있게 한다.

[0047] 이를 시험하기 위해, 본 발명자들은 먼저 디지탈 PCR이 모체 혈장 중 염색체 21로부터의 태반 전사체인 PLAC4 mRNA (문헌 [Lo, YMD. et al. 2007 Nat Med 13, 218-223]) 대립유전자 비율을 정확하고, 이로써 삼염색체 21 및 정배수체 태아를 구별할 수 있는지 여부를 평가하였다. 이러한 접근법을 디지탈 RSA-SNP 방법이라 칭한다. 이어서, 본 발명자들은 디지탈 PCR의 증가된 정확도로 인해 유전자 다형성에 좌우되지 않으면서 태아 염색체 이수성을 검출할 수 있는지 여부를 평가하였다. 본 발명자들은 이를 디지탈 상대적 염색체 투여 (RCD) 분석이라 칭한다. 전자의 접근법은 다형성-의존적이지만 정량 구별에 있어서 보다 낮은 정확도를 필요로 하는 반면, 후자의 접근법은 다형성-독립적이지만 정량 구별에 있어서 보다 높은 정확도를 필요로 한다.

[0049] 1. 디지탈 RNA-SNP

[0050] A. 개관

[0051] 디지탈 PCR은 DNA 샘플 중 두 대립유전자의 벗어난(skewing) 대립유전자 비율의 존재를 검출할 수 있다. 예를 들어, 디지탈 PCR은 종양 DNA 샘플 중 이형접합성 소실 (LOH)을 검출하는데 사용되어 왔다. DNA 샘플 중 두 대립유전자, 즉 A 및 G가 존재하고, A 대립유전자는 LOH가 있는 세포 중에 소실된다고 가정한다. LOH가 종양 샘플 중 50%의 세포에 존재하는 경우, DNA 샘플 중 G:A의 대립유전자 비율은 2:1일 것이다. 그러나, LOH가 종양 샘플에 존재하지 않는 경우, G:A의 대립유전자 비율은 1:1일 것이다.

[0052] 도 1은 디지탈 PCR 실험을 설명하는 흐름도 100이다. 단계 110에서, DNA 샘플을 희석시킨 후, 복수의 웰로 분배한다. 본 발명자들은 일부 혈장 핵산 종이 원래의 샘플에서 이미 상당히 희석되도록 결정했다는 것을 주목한다. 따라서, 필요한 농도로 이미 존재하는 경우, 일부 주형에 대한 희석은 필요하지 않다. 이전의 연구 (예를 들어, 상기 문헌 [Zhou et al 2002])에서는, 특정 "주형 DNA"의 평균 농도가 웰 당 두 주형 중 하나의 대략 0.5 분자가 될 정도로 DNA 샘플을 희석시켰다. 용어 "주형 DNA"는 A 또는 G 대립유전자를 가리키는 것으로 보이고, 이러한 특정 농도에 대한 이론적 근거는 제공되지 않았다는 것을 주목한다.

[0053] 단계 120에서, 각각의 웰에서, PCR 과정은 A 및/또는 G 대립유전자를 동시에 검출하기 위해서 실시된다. 단계 120에서, 각각의 웰 중 마커가 식별된다 (예를 들어, 형광으로) (예를 들어 A, G, A 및 G, 또는 어느 쪽도 아님). LOH의 부재 하에, DNA 샘플 중 A 및 G 대립유전자의 과다는 동일하다 (웰 당 각각 한 카피). 따라서, 웰이 A 대립유전자 및 G 대립유전자에 대해 양성일 확률은 동일할 것이다. 이것은 유사한 수의 웰이 A 또는 G 대립유전자에 대해 양성인 것에 의해 반영될 것이다. 그러나, LOH가 종양 샘플 중 50% 또는 그 보다 많은 세포에 존재하는 경우, G 및 A 대립유전자의 대립유전자 비율은 적어도 2:1일 것이다. 이전의 방법들은 단순히 샘플이 적어도 50% 양성인 것으로 추정하였다. 따라서, 웰이 G 대립유전자에 대해 양성일 확률은 A 대립유전자에 대해 양성일 확률보다 높을 것이다. 그 결과, G 대립유전자에 대해 양성인 웰의 수는 A 대립유전자에 대해 양성인 웰의 수보다 더 많을 것이다.

[0054] 단계 140에서, 디지탈 PCR 결과를 분류하기 위해, 각각의 대립유전자에 대해 양성이나 다른 대립유전자에 대해서는 양성이 아닌 웰의 수를 계수할 것이다. 상기 예에서, A 대립유전자에 대해 양성이나 G 대립유전자에 대해

응성인 웰의 수, 및 G 대립유전자에 대해 양성이나 A 대립유전자에 대해 음성인 웰의 수를 계수한다. 한 실시양태에서, 덜 양성인 웰을 나타내는 대립유전자는 참조 대립유전자로 간주된다.

[0055] 단계 150에서, 정보제공 웰의 총 수는 두 대립유전자 중 어느 하나에 대해 양성인 웰의 수의 합으로 결정된다. 단계 160에서, 더 많은 양성 웰의 대립유전자에 의해 기여되는 정보제공 웰의 비율 (P_r) (파라미터의 한 예)이 계산된다.

[0056] P_r = 더 많은 양성 웰의 대립유전자에 대해서만 양성인 웰의 수/단지 하나의 대립유전자 (A 또는 G)에 대해 양성인 웰의 총 수.

[0057] 다른 실시양태는 하나의 대립유전자를 갖는 모든 웰을 하나 이상의 대립유전자를 갖는 모든 웰로 나눈 것을 사용할 수 있다.

[0058] 단계 170에서, P_r 값이 대립유전자 불균형을 나타내는지 여부가 결정된다. 정확하고 효율적인 것이 바람직하므로, 이 작업은 간단하지 않다. 불균형을 결정하기 위한 한 방법은 베이지안(Bayesian)-유형 우도 방법, 순차 확률비 검정 (SPRT)을 사용한다. SPRT는 데이타가 축적되면서 두 확률 가설을 비교하도록 하는 방법이다. 즉, SPRT는 디지탈 PCR의 결과를 분류하여 벗어난 대립유전자의 존재 또는 부재를 시사하는 통계적 방법이다. SPRT는 분석될 웰의 수를 최소화하여 소정의 통계적 검증력 및 정확도를 획득하도록 하는 이점을 갖는다.

[0059] 예시적 SPRT 분석에서, 실험 결과는 귀무가설 (null hypothesis) 및 대립가설 (alternative hypothesis)에 대해 시험될 것이다. 대립가설은 샘플 중 벗어난 대립유전자 비율이 있는 경우 수용된다. 귀무가설은 샘플 중 벗어난 대립유전자 비율이 없는 경우 수용된다. P_r 값은 귀무가설 또는 대립가설을 수용하기 위해서 두 분할값과 비교될 것이다. 두 가설 모두 수용되지 않는 경우, 샘플은 관찰된 디지탈 PCR 결과가 바람직한 통계적 신뢰도로 샘플을 분류하기에 충분하지 않다는 것을 의미하는 이분류로 표시될 것이다.

[0060] 귀무가설 또는 대립가설을 수용하기 위한 분할값은 전형적으로, 가설에서 언급된 가정 하에 고정된 P_r 값을 기준으로 계산되어 왔다. 귀무가설에서, 샘플은 벗어난 대립유전자 비율을 나타내지 않는 것으로 가정된다. 따라서, 각 웰이 A 및 G 대립유전자에 대해 양성일 확률은 동일할 것이므로, 예상 P_r 값은 1/2일 것이다. 대립가설에서, 예상 P_r 값은 2/3 또는 0.5와 2/3의 대략 중간 정도, 예를 들어 0.585로 여겨진다. 또한, 제한된 수의 실험으로 인해, (.585+3/N)의 상한값 및 (.585-3/N)의 하한값이 선택될 수 있다.

[0062] B. 다운 증후군의 검출

[0063] 본 방법의 한 실시양태에서, 디지탈 SNP는 임신 여성의 혈장으로부터 태아 다운 증후군을 검출하는데 사용된다. 태아/태반 세포에 특이적인 마커를 사용하여, 염색체 21 중 대립유전자의 비율을 측정할 수 있다. 예를 들어, PLAC4 대립유전자의 관찰된 과다표출 정도가 통계적으로 유의한지를 결정하기 위해서, SPRT가 사용된다.

[0064] 하나의 예시적 실시양태에 따라, 디지탈 RNA-SNP는 염색체 21로부터 전사되고 태반에 의해 발현되는 PLAC4 mRNA 에 위치한 A/G SNP의 다형성 대립유전자, rs8130833의 비율에서의 불균형을 결정한다. 이형접합 정배수체 태아에 대해, A 및 G 대립유전자는 태아 게놈에서 동등하게 표출되어야 하는 한편 (1:1 게놈 비율); 삼염색체 21에서, 삼염색체성 염색체 21은 태아 게놈 중 하나의 SNP 대립유전자의 추가 카피와 회합되어 2:1 비율을 나타낼 것이다. 디지탈 PCR 분석의 목적은 분석된 샘플 중 두 PLAC4 대립유전자의 양이 동등한지 또는 다른지를 결정하는 것이다. 따라서, A 및 G PLAC4 대립유전자는 모두 표적 주형이다. 실시간 PCR 검정은 PLAC4 mRNA를 증폭시키도록 설계되었고, 두 SNP 대립유전자는 테크만(TaqMan) 형광 프로브에 의해 구별되었다. 분석 단계의 도식적 설명이 도 2a에 나타나 있다.

[0065] 도 2a는 본 발명의 실시양태에 따른 디지탈 RNA-SNP 방법 200을 설명한다. 단계 210에서, 샘플을 얻는다. 단계 220에서, 핵산 서열, 예를 들어 PLAC4 mRNA를 추출된 RNA 샘플에서 정량화한다. 한 실시양태에서, 이것은 PLAC4 mRNA에 대해 실시간 PCR로 실시된다. 한 측면에서, 이 단계는 조작자에게 표적이 디지탈 PCR 분석의 '영역'에 도달하기 전에 어느 만큼의 희석이 요구되는지에 대한 아이디어를 제공한다.

[0066] 단계 230에서, 샘플을 희석시킨다. 단계 240에서, 희석시킨 샘플의 농도를 측정한다. 희석시킨 샘플의 농도가 약 1 주형/웰 (즉, 참조 또는 비-참조 서열 또는 어느 하나의 대립유전자)인지 확인할 수 있다. 일부 실시양태는 이 측정을 위한 IV절에 기술된 기술을 사용한다. 예를 들어, 본 발명자들은 실시간 PCR 분석을 위해서 희석시킨 샘플을 96 웰로 배분하여 알맞게 희석되었는지를 확인하였다. 희석 농도는 또한 모르는 상태로 남겨두어

이 단계가 제거되도록 할 수 있고, 이는 추후 설명될 것이다.

[0067] 단계 250에서, 디지털 PCR을 어레이의 각 웰 상에서 실시하였다. 예를 들어, 동일하게 희석시킨 샘플을 실시간 PCR 분석을 위해서 384-웰에 배분하였다. PCR 결과로부터, 각 핵산 서열에 대한 아키의 양 및 정보제공 웰의 수를 확인한다. 정보제공 웰은 A 또는 G 대립유전자에 대해서만 양성이나, 둘 모두에 대해서 양성은 아닌 것으로 정의된다. 단계 260에서, 예상 P_i 값을 계산한다. 이러한 단계들은 추후 더욱 자세하게 설명될 것이다. 이 계산은 단계 250에서 측정된 값들로부터 파라미터를 결정하는 것을 포함한다. 예를 들어, 웰 당 실제 평균 주형 농도가 계산될 수 있다.

[0068] 단계 270에서, SPRT 또는 다른 우도-비율 시험은 불균형이 존재하는지 여부를 결정하기 위해서 실시될 수 있다. 정배수체 사례에 대해, 동일한 수의 A-양성 및 G-양성 웰이 예상된다. 그러나, 삼염색체 21 태아로부터의 주형 분자가 분석되는 경우, 하나의 대립유전자만을 함유하는 웰의 수는 다른 대립유전자만을 함유하는 웰의 수보다 많아야 한다. 요컨대, 삼염색체 21에 대한 대립유전자 불균형이 예상된다.

[0069] 상기 언급된 것과 같이, SPRT는 데이타가 축적되면서 두 확률 가설을 비교하도록 하는 베이지안-유형 우도 방법이다. 삼염색체 21 검출을 위한 디지털 PCR 분석에서, 대립가설은 대립유전자 불균형이 존재하는 경우 (즉, 삼염색체 21이 검출되는 경우) 수용되고; 귀무가설은 대립유전자 불균형이 존재하지 않는 경우 (즉, 삼염색체 21이 검출되지 않는 경우) 수용된다. 더 많은 수의 계수의 대립유전자는 잠재적으로 과다표출된 대립유전자로 지정되고, 모든 정보제공 웰 중 그의 비율 (P_i)이 계산될 것이다. SPRT는, P_i이 삼염색체 21 샘플에 대해 예상되는 충분한 정도의 대립유전자 불균형을 나타내는지를 결정하기 위해서 적용된다.

[0070] 조각 연에 있어서, SPRT는 어느 하나의 가설을 수용하거나 또는 거절하기 위한 확률 경계를 정의하기 위해서 구축된 한 쌍의 SPRT 곡선을 갖는 그래프를 사용하여 적용되고 해석될 수 있다. 도 3은 본 발명의 한 실시양태에 따라 다른 종후군을 결정하기 위한 SPRT 곡선을 갖는 그래프를 설명한다. SPRT 곡선은 신뢰성 있게 분류될 수 있는 경우 주어진 정보제공 웰의 총 수 (x-축)에 대해 잠재적으로 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 요구되는 비율 P_i (y-축)을 플롯팅한다. 도 3에 설명된 것과 같이, 상단 곡선은 대립가설을 수용하기 위한 확률 경계를 설정하고; 하단 곡선은 귀무가설을 수용하기 위한 확률 경계를 설정한다.

[0071] 실험적으로 유도된 P_i 값은 어느 하나의 가설을 수용하거나 또는 거절하기 위한 예상 P_i 값과 비교될 것이다. 귀무가설이 수용되는 경우, 샘플은 정배수체 태아를 가진 임신 여성으로부터 수득된 것으로 분류되었다. 대립가설이 수용되는 경우, 샘플은 삼염색체 21 태아를 가진 임신 여성으로부터 수득된 것으로 분류되었다. 별법으로, 주어진 정보제공 계수의 수에 대한 P_i이 질환 분류에 대해 요구되는 통계적 신뢰성 수준에 아직 도달하지 못한 경우, 어느 하나의 가설도 받아들여질 수 없다. 이러한 경우들은 더 많은 데이타가 이용가능하기 전까지는 분류불가능으로 간주된다. 질환 분류가 가능하지 않은 경우, 집계된 데이타가 SPRT에 의해 분류가능해지기 전까지 추가의 384-웰 플레이트를 실시할 수 있다.

[0072] 따라서, SPRT는 주어진 수준의 신뢰성을 위해서 다른 통계적 방법보다 더 적은 양의 서열이 요구된다는 이점을 제공한다. 실용적인 측면에서, SPRT는 요구되는 양의 데이타가 축적되어 불필요한 추가적 분석을 최소화하자마자 어느 하나의 가설의 수용 또는 거절을 가능하게 한다. 이 특징은 일반적으로 이용가능한 주형 분자의 수가 제한적인 경우 낮은 농도로 존재하는 혈장 핵산의 분석과 특히 관련이 있다. 엄격한 분류에 추가하여, 분류는 또한 백분율 정확성을 포함할 수 있다. 예를 들어, 분별값과의 비교로 인한 분류는 샘플이 특정 백분율의 핵산 서열 불균형의 우도를 나타내거나, 또는 동등하게는 측정된 불균형이 특정 백분율 또는 다른 값에 정확하다는 것을 제공할 수 있다.

[0073] 유사한 접근법을 적용하여, 모계 혈장 또는 혈청 중 태아 핵산을 사용하여 돌연변이 또는 유전자 다형성과 관련한 태아의 유전자형을 결정할 수 있다. 태아는 그의 모계로부터 그의 계놈의 절반을 물려받을 것이라는 것을 상기해야 한다. 설명을 위해서, 두 대립유전자 A 및 B가 있는 특정 유전자좌를 고려한다. 모계가 AB 유전자형의 이형접합체인 경우, 태아는 이론적으로는 AA, BB 또는 AB의 유전자형을 가질 수 있다. 태아가 AB 유전자형, 즉 모계와 같은 유전자형을 갖는 경우, 모계 혈장에는 (모계 및 태아 모두로부터의) AB 유전자형의 핵산만 존재할 것이다. 따라서, 핵산 또는 대립유전자 균형이 모계 혈장에서 보여진다. 반면, 태아가 AA 또는 BB의 유전자형을 갖는 경우, 모계 혈장에서 A 또는 B 대립유전자 각각이 과다표출된 대립유전자 불균형이 존재할 것이다. 이러한 생각은 또한 질환-유발성 돌연변이 (예를 들어, 낭포성 섬유증 또는 베타-지중해빈혈 또는 혹수 근위축증을 유발하는 돌연변이)에도 적용가능한데, 각 경우 A는 야생형 대립유전자로, B는 돌연변이 대립유전자로 고

이될 수 있다.

[0075] II. 디지탈 RCD

[0076] 디지탈 RNA-SNP의 단점은 디지탈 RNA-SNP가 분석된 SNP에 대한 이형접합체의 경우에만 적용가능하다는 것이다. 순환성 태아 핵산 분석을 기준으로 한 삼염색체 21 태아 또는 다른 태아 염색체 이수성 (예를 들어, 삼염색체 18, 13 및 성 염색체 이수성)을 검출하기 위한 비침습성 시험이 유전자 다형성의 사용과 무관한 경우 디지탈 RNA-SNP가 이상적이라는 것은 하나의 개선된 점이다. 따라서, 한 실시양태에서, 염색체 양은 참조 염색체, 즉 이 연구에서 염색체 1에 위치한 유전자좌에 상대적인 비-다형성 염색체 21 유전자좌의 디지탈 PCR 분석에 의해 결정된다. 정배수체 태아의 세포에서 염색체 21 대 염색체 1의 비율의 2:2로부터의 변화는 삼염색체 21의 경우와 구분된다. 삼염색체 21 진출을 위한 디지탈 PCR 분석에서, 비교할 두 가설은 염색체 불균형이 존재하지 않는 (즉, 삼염색체 21이 검출되지 않는) 귀무가설 및 염색체 불균형이 존재하는 (즉, 삼염색체 21이 검출되는) 대립가설일 것이다.

[0077] 이 접근법은 다른 염색체 이수성, 예를 들어 삼염색체 18에서의 염색체 18, 삼염색체 13에서의 염색체 13, 터너 증후군에서의 염색체 X와 연관된 다른 염색체에 일반화될 수 있다. 또한, 염색체 1과는 멀개로, 관련 이수성과 연관되지 않은 다른 염색체도 또한 참조 염색체로 사용될 수 있다. 유사한 접근법이 또한 부분적으로, 암에서 통상적으로 결실되는 염색체 대 참조 염색체의 비율 변화를 분석함으로써 암 검출에 적용될 수 있다. 전자의 예로는 결장직장암에서의 염색체 5q, 폐암에서의 염색체 3p 및 비인두 암종에서의 염색체 9p가 포함된다. 도 2b는 서열 불균형으로 나타나는 일부 통상적인 암-관련 염색체 이상을 열거한다.

[0078] 도 2a는 또한 본 발명의 한 실시양태에 따른 디지탈 RCD 방법 205를 설명한다. 단계 220-230에 대한 한 실시양태에서, 추출된 DNA을, 예를 들어, 나노드롭(Nanodrop) 기술로 정량화하고, 웰 당 염색체 21 또는 표준화 염색체 (예컨대, 염색체 1)로부터의 약 하나의 표적 주형의 농도로 희석시킨다. 단계 240의 한 실시양태에서, 384-웰 플레이트 중 헤르만 프로브를 모두 사용하는 디지탈 RCD 분석을 진행하기 전에 약 37% 수준의 웰이 음성인지 여부를 확인하기 위해서, 단지 96-웰 포맷에서 염색체 1 프로브를 사용하는 검정에 의해 희석시킨 DNA 샘플을 분석함으로서 확인할 수 있다. 37%의 유의성은 IV절에서 추후 기술될 것이다.

[0079] 단계 240의 시험 및 단계 250의 결과는 한 쌍의 헤르만 프로브에 의해 구별되는 파라로그(paralogous) 서열 변이에 의해 식별되는 두 염색체 모두에 존재하는 파라로그 서열을 증폭시키도록 고안된 실시간 PCR 검정으로 수행될 수 있다 (문헌 [Deutsch, S. et al. 2004 J Med Genet 41, 908-915]). 이러한 맥락에서, 정보제공 웰은 염색체 21 또는 염색체 1 유전자좌에 대해 양성인, 그러나 둘 모두에 대해서 양성은 아닌 웰로서 정의된다. 정배수체 태아에 대해, 어느 하나의 유전자좌에 대해 양성인 정보제공 웰의 수는 대략 동등해야 한다. 삼염색체 21 태아에 대해, 염색체 1보다 염색체 21에 대해 양성인 웰의 과다표출이 있어야 한다. 과다표출의 정확한 비율은 하기 절에 기술되어 있다.

[0081] III. 태아 서열의 존재율

[0082] 상기 기술된 방법 200 및 205의 실시양태의 단점은 태아 특이적 마커가 요구된다는 것이다. 따라서, 본 발명의 한 실시양태에서, 비-태아 특이적 마커가 사용된다. 비-태아 특이적 마커를 사용하기 위해서, 본 발명의 실시양태는 모체 혈장 (즉, 생물학적 샘플) 중 태아 DNA 분획 농도를 측정한다. 이러한 정보와 함께, 보다 유용한 P_1 값이 하기와 같이 계산될 수 있다.

[0083] 모체 혈장 중 태아 DNA의 낮은 분획 백분율로도, 삼염색체 21 태아는 모체 혈장으로 방출된 태아 DNA의 게놈-등가물 (GE) 당 염색체 21 서열의 추가적 양에 기여할 것이다. 예를 들어, 태아에 의해 기여된 5 GE/mL DNA와 50 GE/mL의 총 DNA (즉, 10% 태아 DNA 분획 농도)를 함유하는 정배수체 임신으로부터의 모체 혈장 샘플은 모체 혈장 (밀리리터) 당 염색체 21 서열의 총 100개의 카피 (90개의 모체 카피 + 10개의 태아 카피)를 함유해야 한다. 삼염색체 21 임신에 대해, 태아 GE는 각각 3개 카피의 염색체 21에 기여하여, 모체 혈장 중 염색체 21 서열의 총 105개의 카피/mL (90개의 모체 카피 + 15개의 태아 카피)가 되도록 한다. 따라서, 10% 태아 DNA 농도에서, 삼염색체 임신의 모체 혈장 중 염색체 21 유래 서열의 양은 정배수체 경우의 1.05배일 것이다. 따라서, 분석적 접근법이 이러한 적은 정도의 경량 차이를 결정하도록 개발될 수 있다면, 삼염색체 21 태아의 비침습성 출생 전 진단에 대한 다형성-독립적 시험이 가능할 것이다.

[0084] 따라서, 과다표출의 정도는 분석된 DNA 샘플 중 태아 DNA 분획 농도에 좌우될 것이다. 예를 들어, 태반 DNA를 분석하는 경우, 태아 게놈 중 이론적 RCD 비율은 2:2, 즉 1.5-배 차이어야 한다. 그러나, 상기 기술된 것과 같이, 이론적 RCD 비율은 10% 태아 DNA를 함유하는 모체 혈장 샘플을 분석하는 경우 1.05로 감소할 것이다.

실험적으로 유도된 P_n은 단지 염색체 21 유전자좌에 양성인 핼의 수를 총 검출핼 핼의 수로 나누어 계산된다. 실험적으로 유도된 P_n은 계산된 P_n 및 이론적 RCD 비율과 함께 SPRT 분석에 적용된다.

[0085] 도 4는 본 발명의 실시양태에 따라 태아 핵산의 백분율을 사용하여 질환 상태를 결정하는 방법 400을 나타낸다. 단계 410에서, 태아 물질의 분획 백분율이 측정된다. 한 실시양태에서, 분획 백분율은 비-태아-특이적 마커 (즉, 모계 및 태아 모두에 존재하는 유전자 서열)와 관련하여 태아-특이적 마커 (예를 들어, Y-염색체, 유전자 다형성 마커 (예컨대, SNP), 태반 후성 기호 (placental epigenetic signatures))의 양을 측정하여 결정된다. 실시간 PCR, 디지털 PCR, 서열분석 반응 (대규모 병렬 게놈 서열분석을 포함) 또는 임의의 다른 공지 방법에 의해 실질적 측정을 수행할 수 있다. 한 측면에서, 이 측정에 대해 잠재적으로 대립유전자 불균형이 있을 수 있는 유전자 표적을 사용하지 않는 것이 바람직하다.

[0086] 단계 420에서, 샘플을 희석시키고, 희석시킨 샘플을 핼에 넣고, 각 핼에서 반응을 측정하는 것을 비롯한 디지털 PCR 또는 다른 측정 방법이 실시된다. 단계 430에서, PCR 결과는 상이한 참조 핵산 서열 (예컨대, 염색체 또는 대립유전자)의 아이를 식별하기 위해서 사용된다. 단계 440에서, 과다표출된 서열의 실제 비율 (P_n)이 계산된다. 단계 450에서, 질환 상태를 결정하기 위한 문별값은 샘플 중 태아 물질의 백분율을 사용하여 계산된다. 단계 460에서, 실제 P_n 및 문별값으로부터, 돌균형이 존재하는지 여부가 결정된다.

[0087] 한 실시양태에서, 참조 핵산 서열의 분획 백분율은 디지털 RNA-SNP 방법에 혼입된다. 따라서, 암 세포로 인한 LOH를 연구하는 경우, 50% 미만 암 세포를 갖는 종양 샘플로 실시할 수 있다. 이것은 또한 50% 초과 암 세포를 갖는 샘플에 대해서도 사용되어, 더욱 정확한 P_n을 수득함으로써 부정확한 진단으로 이끌 수 있는 잘못된 양성의 수를 감소시킬 수 있다. 또다른 실시양태에서, 태아 핵산 백분율은 디지털 PCR 방법에 혼입되어, 태아가 모체 혈장 핵산 분석으로부터 부모의 유전자 돌연변이 (예를 들어, 낭포성 섬유증 또는 베타-지중해빈혈 또는 희수 근위축증을 유발하는 돌연변이) 또는 다형성을 물려받는지 여부를 결정한다.

[0089] IV. 핼 당 평균 농도의 혼입

[0090] 이전 방법 (예를 들어, 상기 문헌 [Zhou, X. et al., 2002])의 또다른 단점은 핼 당 주형의 평균 농도 (m)가 핼 당 1로 요구된다는 것이다. 정확한 농도를 얻는 것이 어렵다는 것을 감안할 때, 이것은 부정확해질 수 있다. 또한, 핼 당 1 주형의 정확한 농도로도, 이런 방법은 핼 중 주형의 통계적 분포를 간과하였다. 이전 방법에서, 즉, 기존 알고리즘에서, 대립가설을 수용하기 위한 예상 P_n 값은 대립유전자 비율로 가정되어, 핼 당 주형 DNA의 평균 농도와 무관하였다.

[0091] 그러나, 희석시킨 샘플에서 주형의 자연적인 통계적 편차로 인해, 정확하게 핼 당 1 주형은 아닐 것이다. 본 발명의 실시양태는 하나 이상의 서열의 평균 농도를 측정하여, 분별값, 즉 예상 P_n을 계산하는데 사용된다. 한 측면에서, 이러한 계산은 상이한 핵산 서열을 함유하는 핼의 확률을 측정하여 예상 P_n을 결정하기 위해서 사용되는 통계적 분포를 포함한다.

[0092] 한 실시양태에서, 평균 농도는 일배로 DNA 샘플 중 더 낮은 농도의 핵산 서열인 한 참조 핵산 서열로 고려된다. 돌균형이 없는 샘플의 경우, 샘플 중 두 서열의 농도는 동일할 것이고, 어느 하나는 참조 대립유전자로 간주될 수 있다. 예를 들어 LOH가 있는 샘플의 경우, 암 세포에서 결실된 대립유전자는 참조 대립유전자로 간주될 것이다. 참조 대립유전자의 평균 농도는 m_r로 표시될 것이다. 또다른 실시양태에서, 더 높은 농도의 서열이 참조 서열로 간주될 수 있다.

[0094] A. 디지털-SNP, SPRT 및 디지털 PCR을 사용한 예

[0095] 도 5는 본 발명의 한 실시양태에 따라 평균 주형 농도를 사용하여 질환 상태를 결정하는 방법 500을 보여준다. 단계 510에서, 상이한 서열의 양을 측정한다. 이것은, 예를 들어 상기 설명된 것과 같은 디지털 PCR 실험에서 마커를 계수하여 실시할 수 있다. 그러나, 이것은 중폭 단계를 포함하지 않거나 또는 형광 마커를 사용하지 않지만 다른 성질, 예컨대 질량, 특정 광학적 성질 또는 엄기-쌍 성질과 같은 물성을 사용할 수 있는 다른 방법에 의해서도 실시할 수 있다.

[0096] 단계 520에서, 과다표출된 서열의 실제 비율이 결정된다. 이것은 상기 기술된 것과 같이 단지 그 서열만 나타내는 핼의 수를 확하고, 정보제공 핼의 수로 나누어 실시될 수 있다. 단계 530에서, 하나 이상의 서열의 평균 농도가 측정된다 (참조 서열). 한 실시양태에서, 참조 서열은 과다표출된 서열이다. 또다른 실시양태에서, 참

조 서열은 과소표출된(underrepresented) 서열이다. 측정은 디지탈 PCR 실험에서 참조 서열에 대해 음성인 웰의 수를 계수하여 실시할 수 있다. 음성 웰의 비율과 평균 주형 농도 사이의 관계는 다음 소절에 기술된 것과 같은 푸아송 (Poisson) 분포에 의해 설명된다.

[0097] 단계 540에서, 상이한 서열에 대해 양성인 웰의 예상 양이, 예를 들어 푸아송 분포를 사용하여 계산된다. 예상 양은 웰 당 서열, 웰 당 평균 서열, 서열을 함유하는 웰의 수 또는 임의의 다른 적합한 양의 확률일 수 있다. 단계 550에서, 예상 P_i은 예상 양으로부터 계산된다. 단계 560에서, 분별값은, 예를 들어 SPRT를 사용하여 예상 P_i로부터 계산된다. 단계 570에서, 핵산 서열 불균형의 분류가 결정된다. 방법 500의 특정 측면이 이제 기술된다.

[0099] 1. 서열 예상 양의 결정

[0100] 웰 (반응 또는 반응 혼합물) 당 평균 농도가 단계 530으로부터 알려지면, 서열을 나타내는 웰의 예상 수가 단계 540에서 계산될 수 있다. 이 양은 %, 분수 값, 또는 정수 값으로 표현될 수 있다. 설명을 위해 특정 예를 사용하면서, 웰 당 참조 주형의 평균 농도 (m_r)는 웰 당 0.5이고, PLAG1 SNP, rs8130833에서 삼염색체 21 태아의 유전자형은 AGG인 것으로 가정한다. 따라서, 참조 주형은 A 대립유전자일 것이고, 과다표출된 주형은 G 대립유전자일 것이다.

[0101] 한 실시양태에서, 푸아송 분포는 디지탈 PCR과 같은 측정 과정의 웰의 반응 혼합물 중 A 대립유전자의 분포로 가정한다. 다른 실시양태에서는, 이항 분포와 같은 다른 분포 함수가 사용된다.

[0102] 푸아송 방정식은 $P(n)=\frac{m^n e^{-m}}{n!}$ 이고, 여기서 n은 웰 당 주형 분자의 수이고; P(n)은 특정 웰 중 n 주형 분자의 확률이고; m은 특정 디지탈 PCR 실험에서 한 웰 중 주형 분자의 평균 수이다.

[0103] 따라서, 0.5의 평균 A-대립유전자 농도에서 A 대립유전자의 임의의 분자를 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

[0104] $$P(0)=\frac{0.5^0 e^{-0.5}}{0!}=e^{-0.5}=0.6065$$

[0105] 따라서, 하나 이상의 A 대립유전자의 분자를 함유하는 임의의 웰의 확률은: 1 - 0.6065 = 0.3935이다. 그러므로, 약 39%의 웰이 하나 이상의 A 대립유전자의 분자를 함유할 것으로 예상될 것이다.

[0106] 비-참조 핵산 서열의 경우, 삼염색체 21 태아의 각 세포에 대해서, A 대 G의 게놈 비율은 1:2일 것이다. 추출된 RNA 또는 DNA 샘플 중 A 대 G 비율이 변화하지 않는 것으로 가정하면, 웰 당 G 대립유전자의 평균 농도는 A 대립유전자의 평균 농도의 2배, 즉 2 × 0.5 = 1일 것이다.

[0107] 따라서, 평균 G-대립유전자 농도인 1에서 G 대립유전자의 임의의 분자를 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

[0108] $$P(0)=\frac{1^0 e^{-1}}{0!}=e^{-1}=0.3679$$

[0109] 따라서, 하나 이상의 G 대립유전자 분자를 함유하는 임의의 웰의 확률은: 1 - 0.3679 = 0.6321일 것이다. 그러므로, 약 63%의 웰이 하나 이상의 G 대립유전자 분자를 함유하는 것으로 예상될 것이다.

[0110] 2. 과다표출된 서열 비율의 결정

[0111] 예상 양이 계산된 후, 과다표출된 핵산 서열의 비율이 결정될 수 있다. A 대립유전자 및 G 대립유전자로 웰을 채우는 것이 독립적이라고 가정하면, 두 대립유전자를 모두 함유하는 웰의 확률은 0.3935 × 0.6321 = 0.2487일 것이다. 그러므로, 약 25%의 웰이 두 대립유전자를 모두 함유하는 것으로 예상될 것이다.

[0112] A 대립유전자는 함유하의 G 대립유전자는 함유하지 않는 것으로 예상되는 웰의 비율은 하나 이상의 A 대립유전자를 함유하는 웰의 수에서 A 및 G 대립유전자를 모두 함유하는 웰의 수를 뺀 것이다: 0.3935 - 0.2487 = 0.1448. 유사하게, G 대립유전자는 함유하되 A 대립유전자는 함유하지 않는 것으로 예상되는 웰의 비율은 0.6321 - 0.2487 = 0.3834일 것이다. 정보제공 웰은 A 대립유전자 또는 G 대립유전자에 대해 양성이나, 둘 모두에 대해서는 양성이 아닌 웰로서 정의된다.

[0113] 따라서, 디지탈 RNA-SNP 분석에서 G 대립유전자에 상대적으로 A 대립유전자를 함유하는 웰의 예상 비율은 0.1448/0.3834이다. 즉, 단지 G 대립유전자에 대해서만 양성인 웰의 비율은 단지 A 대립유전자에 대해서만 양성인 웰의 2.65배이다. 이것은 과다표출된 대립유전자가 다른 대립유전자의 2배인 때의 계능 비율과 대조적이다.

[0114] SPRT 분석에 대해, 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 비율 (P_1)은 SPRT 곡선을 사용하여 계산되고 해석된다. 현 예에서, 정보제공 웰의 비율은 0.1448 + 0.3834 = 0.5282일 것이다. 따라서, m_c 0.5에서 삼염색체 21 경우의 예상 P_1은 0.3834/0.5282 = 0.73이다.

[0115] 평균 주형 농도 (m)는 푸아송 방정식의 주요 파라미터이므로, P_1은 m에 따라변할 것이다. 도 6은 본 방법의 실시양태에 따라 웰 당 평균 참조 주형 농도 (m_c)로 표현된 주형 농도의 범위에 대한 삼염색체 21 샘플의 예상 디지탈 RNA-SNP 대립유전자 비율 및 P_1을 표로 만든 표 600을 나타낸다. 표 600은 일련의 웰 당 평균 참조 주형 농도 (m_c)에 대한 예상 대립유전자 비율 및 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 비율 (P_1)을 나타낸다.

[0116] 예상 P_1 값은 비-선형 방식으로 웰 당 참조 대립유전자의 평균 농도 (m_c)에 따라 변한다. 표 600에 제시된 것과 같이, 대립가설을 수용하기 위한 예상 P_1 값은 m_c과 함께 증가할 것이다. 귀무가설을 수용하기 위한 예상 P_1 값이 0.5로 고정되어 있으므로, 대립유전자 불균형이 있는 샘플 및 대립유전자 불균형이 없는 샘플은 m_c이 증가할 경우, P_1 값의 측면에서 보다 광범위하게 나뉘어질 것이다. 다른 실시양태에서는 귀무가설을 수용하기 위한 값이 0.5가 아닐 수 있다는 것을 주목한다. 이것은 정상 비율이 1:1과 다른 경우, 예를 들어, 5:3인 경우 일어날 수 있으며, 따라서 비율이 5:2으로부터 벗어나는 경우 불균형이 발생할 것이다. 두 상이한 핵산 서열 양의 차이는 개별 경우에 따라 결정될 것이다.

[0117] 그러나, 이전 방법 (예를 들어, 상기 문헌 [Zhou, W. et al., 2002])은 LOH 샘플에 대해 고정된 예상 P_1 값을 사용하였으므로, LOH가 있는 샘플에 대한 P_1 값을 과소평가하였다 (대립가설 수용). 과소평가의 정도는 m_c과 함께 증가할 것이다. 즉, DNA 샘플 중 참조 대립유전자의 평균 농도가 더 높을수록, 이전 방법은 더욱 부정확해질 것이다. 대립가설을 수용하기 위한 P_1의 과소평가는 귀무가설 및 대립가설 모두를 수용하기 위한 분별값의 계산을 부정확하게 할 것이다.

[0119] 3. 예상 P_1을 기준으로 한 분별값의 계산

[0120] SPRT를 사용하는 실시양태에 대해, 문헌 [El Karoui at al., (2006)]으로부터 SPRT 곡선의 상한 및 하한을 계산하기 위한 방정식을 사용할 수 있다. 또한, 귀무가설 또는 대립가설을 수용하기 위해서 바람직한 통계적 신뢰성의 수준은 방정식에서 임계값 우도 비율(threshold likelihood ratio)을 조정하면서 변할 수 있다. 이 연구에서, 임계값 우도 비율 8이 양 검출의 맥락에서 대립유전자 불균형이 있는 샘플과 대립유전자 불균형이 없는 샘플을 구별하는 만족할만한 성능을 제공하는 것으로 나타났기 때문에, 상기 값 8이 사용된다. 따라서, 한 실시양태에서, SPRT 곡선의 상한 및 하한을 계산하기 위한 방정식은 다음과 같다:

[0121] 상한 = $[(\ln 8)/N - \ln \delta]/\ln \gamma$

[0122] 하한 = $[(\ln 1/8)/N - \ln \delta]/\ln \gamma$

[0123] 여기서, $\delta = (1 - \Theta_1)/(1 - \Theta_0)$,

[0124] $\gamma = -(\Theta_1(1 - \Theta_0)/\Theta_0(1 - \Theta_1)$,

[0125] Θ_0 = 귀무가설이 사실인 경우 비-참조 대립유전자를 함유하는 정보제공 웰의 비율 = 0.5 (하기 참조),

[0126] Θ_1 = 대립가설이 사실인 경우 비-참조 (즉, 과다표출된) 대립유전자를 함유하는 정보제공 웰의 비율,

[0127] N = 정보제공 웰의 수 = 어느 하나의 대립유전자에만 양성인 웰의 수,

[0128] (ln은 자연 로그, 즉 $\log_e$를 나타내는 수학 기호임)

[0129] 귀무가설을 수용하기 위한 Θ_0의 결정을 위해서, 샘플은 정배수체 태아를 갖는 임신 여성으로부터 얻어지는 것으로 가정된다. 이러한 가정 하에, 어느 하나의 주형에 양성인 웰의 예상 수는 1:1일 것이므로, 비-참조 대립유전자를 함유하는 정보제공 웰의 예상 비율은 0.5일 것이다.

[0130] 대립가설을 수용하기 위한 Θ_1의 결정을 위해서, 샘플은 삼염색체 21 태아를 갖는 임신 여성으로부터 얻어지는 것으로 가정된다. 디지털 RNA-SNP 분석을 위한 삼염색체 21 경우의 예상 P_i의 계산은 표 600에 상술되어 있다. 따라서, 디지털 RNA-SNP 분석을 위한 Θ_1은 표 600의 마지막 열에 제시된 데이타를 참조한다.

[0132] 4. 평균 농도의 측정

[0133] m_i의 측정은 당업자에게 공지되거나 또는 공지될 다양한 메커니즘에 의해 실시될 것이다. 한 실시양태에서, m_i 값은 디지탈 PCR 분석의 실행 과정 중 결정된다. m_i 값과 참조 대립유전자에 대해 양성인 웰의 총 수 사이의 관계는 분포 (예를 들어, 푸아송 분포)에 의해 좌우될 수 있으므로, m_i은 다음 식을 사용하여 참조 대립유전자에 대해 양성인 웰의 수로부터 계산될 수 있다:

[0134] m_i = - ln (1 - 참조 대립유전자에 대해 양성인 웰의 비율).

[0135] ln은 자연 로그, 즉 $\log_e$라는 것에 주목한다. 이 접근법은 디지털 PCR 실험을 위해 사용되는 DNA 샘플 중 m_i의 적절하고 정확한 평가를 제공한다.

[0136] 이 방법은 바람직한 농도를 얻기 위해서 사용될 수 있다. 예를 들어, 샘플의 추출된 핵산은 방법 200의 단계 240에서 실시한 것과 같이 반응 웰 당 하나의 주형 분자와 같은 특정 농도로 희석될 수 있다. 푸아송 분포를 사용하는 한 실시양태에서, 주형이 없는 웰의 예상 비율은 e^{-m}으로 계산될 수 있고, 여기서 m은 웰 당 주형 분자의 평균 농도이다. 예를 들어, 웰 당 하나의 주형 분자의 평균 농도에서, 주형 분자가 없는 웰의 예상 비율은 e^{-1}, 즉 0.37 (37%)로 주어진다. 나머지 63%의 웰은 하나 이상의 주형 분자를 함유할 것이다. 전형적으로, 디지탈 PCR 실험에서 양성 웰 및 정보제공 웰의 수가 계수될 것이다. 정보제공 웰의 정의 및 디지탈 PCR 데이타가 해석되는 방식은 용도에 좌우된다.

[0137] 다른 실시양태에서, 웰 당 평균 농도 m_i은 또다른 정량화 방법, 예를 들어 정량 분광법을 사용하는 정량 실시간 PCR, 완-정량 경쟁적 PCR, 실-경쟁적 PCR 등에 의해 측정된다.

[0139] B. 디지탈 RCD

[0140] 평균 농도를 사용하는 디지탈 RCD를 상기 기재된 디지탈 SNP 방법과 유사한 방식으로 수행할 수 있다. 참조 염색체 (비-염색체 21) 마커, 염색체 21 마커, 및 두 마커 모두에 대해 양성인 웰의 수를 디지탈 PCR에 의해 측정할 수 있다. 디지탈 SNP 분석에 대한 웰 당 참조 마커의 평균 농도 (m_i)의 계산에서와 같은 푸아송 확률 함수에 따라, 염색체 21 마커의 양성에 상관없이, 웰 당 참조 마커의 평균 농도 (m_i)를 참조 마커에 대해 음성인 웰의 총 수로부터 계산할 수 있다.

[0141] 이후, 정배수체 또는 삼염색체 21 태아를 갖는 임신 여성으로부터 얻어진 것과 같은 혈장 샘플을 분류하기 위해 SPRT 분석을 사용할 수 있다. 태아가 정배수체인 경우 귀무가설이 수용될 것이다. 이 시나리오에서, 참조 마커 및 염색체 21 마커에 대해 양성인 웰의 예상 비율은 1:1일 것이며, 이에 따라, 염색체 21 마커에 대해 양성 신호를 갖는 정보제공 웰의 예상 비율은 0.5일 것이다. 태아가 염색체 21에 대해 삼염색체인 경우 대립가설이 수용될 것이다. 이 시나리오에서, 샘플 DNA가 전적으로 태아로부터 유래된 경우, 각 웰에서의 염색체 21 마커의 평균 농도는 참조 마커의 평균 농도 (m_i)의 3/2배일 것이다.

[0142] 디지탈 RCD를 사용하여 태아-특이적 마커, 예를 들어 태반 후성 기호의 검출을 통해 염색체 양을 측정할 수 있지만 (문헌 [Chim, SSC, et al, 2005 Proc Natl Acad Sci USA 102, 14753-14758]), 디지탈 RCD 분석의 실시양태는 비-태아-특이적 마커를 사용한다. 따라서, 비-태아 특이적 마커를 사용하는 경우 태아 활성의 백분율을 측정하는 추가의 단계가 나타날 것이다. 따라서, 웰 당 염색체 21 마커의 평균 농도는 샘플 중 태아 DNA의 비율에 의존적일 것이며, m_i {(200% + 태아 DNA 백분율)/200%}을 사용하여 계산할 수 있다.

[0143] 예시용 특정 실시예를 다시 사용하여, 웰 당 참조 주형인 염색체 1의 평균 농도 (m_1)는 0.5인 것으로 가정하여,

50%의 DNA는 태아로부터 유래된 것으로 가정하고, 샘플 중 50%의 DNA는 모체로부터 유래된다.

[0144] 따라서, 푸아송 분포를 사용하여, 평균 농도가 웰 당 0.5인 경우 염색체 1 유전자좌의 어떠한 분자도 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

$$P(0) = \frac{0.5^0 e^{-0.5}}{0!} = e^{-0.5} = 0.6065$$

[0145]

[0146] 따라서, 염색체 1 유전자좌 중 하나 이상의 분자를 함유하는 임의의 웰의 확률은 1 - 0.6065 = 0.3935일 것이다. 따라서, 대략 39%의 웰이 유전자좌 중 하나 이상의 분자를 함유하는 것으로 예상될 것이다.

[0147] 이러한 삼염색체 21 태아의 각각의 세포에 대하여, 염색체 21 대 염색체 1의 세는 비율은 3:2일 것이다. DNA 샘플에서 염색체 21과 염색체 1 사이의 비율은 태아 DNA 분획 농도 (태아 DNA %)에 의존적일 것이며, 3×태아 DNA % + 2(1 - 태아 DNA %) : 2×태아 DNA % + 2×(1 - 태아 DNA %)일 것이다. 따라서, 이 경우에서, 태아 DNA 분획 농도가 50%인 경우, 비는 (3×50% + 2×50%) / (2×50% + 2×50%) = 1.25일 것이다. 디지탈 SNP 방법이 태아 특이적 마커를 사용하지 않은 경우, 이러한 계산을 사용하여 비-종조 서열의 평균 농도를 계산할 수도 있다.

[0148] 따라서, 웰 당 염색체 1 유전자좌의 평균 농도가 0.5인 경우, 웰 당 염색체 21 유전자좌의 평균 농도는 1.25×0.5 = 0.625이다. 따라서, 평균 농도가 웰 당 0.625인 경우 염색체 21 유전자좌의 어떠한 분자도 함유하지 않는 임의의 웰의 확률은 하기와 같을 것이다:

$$P(0) = \frac{0.625^0 e^{-0.625}}{0!} = e^{-0.625} = 0.5353$$

[0149]

[0150] 따라서, 염색체 21 유전자좌 중 하나 이상의 분자를 함유하는 임의의 웰의 확률은 1 - 0.5353 = 0.4647일 것이다. 따라서, 대략 46%의 웰이 유전자좌 중 하나 이상의 분자를 함유하는 것으로 예상될 것이다. 유전자좌 중 하나로의 웰의 충전이 독립적이라고 가정하면, 유전자좌 모두를 함유하는 웰의 확률은 0.3935×0.4647 = 0.1829일 것이다. 따라서, 대략 18%의 웰이 두 유전자좌 모두를 함유하는 것으로 예상될 것이다.

[0151] 염색체 1 유전자좌를 함유하되 염색체 21 유전자좌는 함유하지 않는 것으로 예상되는 웰의 비율은, 하나 이상의 염색체 1 유전자좌를 함유하는 웰의 수에서 유전자좌 모두를 함유하는 웰의 수를 차감한 것, 즉 0.3935 - 0.1829 = 0.2106일 것이다. 유사하게, 염색체 21 유전자좌를 함유하되 유전자좌 모두는 함유하지 않는 것으로 예상되는 웰의 비율은 0.4647 - 0.1829 = 0.2818일 것이다. 정보제공 웰은 염색체 1 유전자좌 또는 염색체 21 유전자좌 중 하나에 대해 양성이되 둘 모두에 대해서는 아닌 웰로서 정의된다.

[0152] 따라서, 디지탈 RCD 분석에서의 염색체 21 대 염색체 1 예상 비율은 0.2818/0.2106 = 1.34이다. 바꾸어 말하면, 염색체 21 유전자좌에 대해서만 양성인 웰의 비율은 염색체 1 유전자좌에 대해서만 양성인 웰의 비율의 1.34배이다. 이는 DNA 샘플에서의 1.25의 비율과는 대조적이다.

[0153] SPRT 분석에 대하여, 염색체 21 유전자좌에 대해 양성인 정보제공 웰의 비율 (P_1)을 SPRT 곡선을 사용하여 계산 및 해석하는 것이 필요할 것이다. 현재 예에서, 정보제공 웰의 비율은 0.2106 + 0.2818 = 0.4924일 것이다. 따라서, 0.5의 m_0에서 50% 태아 DNA를 갖는 삼염색체 21 경우의 예상 P_1은 0.2818/0.4924 = 0.57이다.

[0154] 평균 주행 농도 (m)가 푸아송 방정식에서의 중요 파라미터이기 때문에, P_1은 m에 의해 변할 것이다. 도 7은 본 방법의 실시양태에 따른 웰 당 평균 참조 주형 농도 (m_0)로서 표현된 주형 농도의 범위에서 삼염색체 21 샘플 중 10%, 25%, 50% 및 100%의 태아 DNA 분획 농도에 대한 예상 P_1을 표로 만든 표 700을 나타낸다. 디지탈 RCD 분석에 대한 삼염색체 21 경우의 예상 P_1에 대한 계산을 표 700에서 상술한다. 따라서, 표 700에서 상응하는 예상 P_1 값을 나타내는 열로부터 여러 태아 DNA 분획 농도를 갖는 샘플의 디지탈 RCD 분석에 대한 G_1을 얻을 수 있다.

[0156] C. 결과

[0157] 1. 상이한 m_0의 비교

[0158] 대립유전자 또는 염색체 불균형의 이론값 (대안 계놈에서와 같음)과 실험적으로 예상되는 정도 사이의 차이, 및 m_c 값 범위에 대한 추자의 측정을 위한 계산값에 대한 근거를 표 600 및 700에 나타낸다. 삼염색체 21 샘플의 디지탈 RNA-SNP 분석에서, m_c = 0.5인 경우, 참조 대립유전자만을 함유하는 웰에 대한 과다표출된 대립유전자만을 함유하는 웰, 즉 디지탈 RNA-SNP 비율은 2.65이다 (표 600). 100% 태아 DNA로 구성된 표본의 디지탈 RCD 분석에서, m_c = 0.5인 경우, 염색체 1 유전자좌에 대해 단독으로 양성인 웰에 대한 염색체 21 유전자좌에 대해 단독으로 양성인 웰, 즉 디지탈 RCD 비율은 1.7이다 (표 700) (P_1 = 0.63, 따라서 디지탈 RCD 비율은 0.63/(1 - 0.63) = 1.7임). 태아 DNA 분획 농도가 감소하기 때문에, 디지탈 RCD 비율은 동일한 m_c에 대해 감소한다 (표 700).

[0159] 표 600 및 700에서 나타낸 것과 같이, 대립유전자 또는 염색체 과다표출 정도는 m_c에 따라 증가한다. 그러나, 정보제공 웰의 백분율은 m_c = 0.5 근처에서 그의 최대에 도달하며, m_c이 추가로 증가함에 따라 점차 감소한다. 실제로, 표본에 대한 주형 분자의 양이 제한되지 않는 경우 정보제공 웰의 비율의 감소는 분석된 웰의 총 수의 증가에 의해 상쇄될 수 있지만, 추가의 웰은 시약 비용의 증가를 요구할 것이다. 따라서, 최적 디지탈 PCR 수행은 주형 농도와 샘플 당 시험된 웰의 총 수 간에 상호 의존적이다.

[0161] 2. SPRT 곡선을 사용하는 예

[0162] 상기 논의된 것과 같이, 디지탈 PCR 실험에 대한 예상 대립유전자 또는 염색체 불균형 정도는 반응 혼합물 (예를 들어, 웰) 당 실제 주형 농도에 의존적이다. 본 발명자들은 참조 대립유전자를 기준으로 한 주형 농도, 즉 웰 당 평균 참조 주형 농도 (m_c)를 기재한다. 상기 방정식에서 나타낸 것과 같이, 예상 P_1을 사용하여 상부 및 하부 SPRT 곡선의 플롯팅을 결정할 수 있다. 예상 P_1이 또한 m_c 값에 의존적이기 때문에, SPRT 곡선의 플롯팅은 본질적으로 m_c 값에 의존적일 것이다. 따라서, 실제로, 디지탈 PCR 데이타 세트의 실제 m_c에 관련된 한 세트의 SPRT 곡선이 특정 실행으로부터 P_1의 해석을 위해 사용될 필요가 있다.

[0163] 도 8은 본 발명의 실시양태에 따른 디지탈 RNA-SNP 분석에 있어서 0.1, 0.5 및 1.0의 m_c 값에 대한 SPRT 곡선에서의 차이의 정도를 예시하는 플롯 800을 나타낸다. 각 세트의 디지탈 PCR 데이타는 특정 실행의 정확한 m_c 값에 관련된 특정 곡선으로 해석되어야 한다. 디지탈 RNA-SNP 및 RCD 방법에 대한 예상 대립유전자 또는 염색체 불균형 정도가 상이하기 때문에 (전자: 2:1, 후자: 3:2), 2개 디지탈 PCR 시스템에 대하여 상이한 세트의 SPRT 곡선이 필요하다는 것을 주목한다. 실험적으로 유도된 P_1을 디지탈 PCR 실행의 상응하는 m_c에 의해 선택된 관련 SPRT 곡선을 이용하여 해석한다. 이는 고정된 세트의 곡선이 사용된 경우 디지탈 PCR에 의한 LOH의 분자 검출을 위해 앞서 보고된 SPRT를 사용한 것과는 대조적이다.

[0164] SPRT를 사용하여 디지탈 PCR 데이타를 해석하기 위한 실제적 방법을 가상 디지탈 RNA-SNP 실행을 사용하여 하기 예시한다. 각 경우의 디지탈 RNA-SNP 분석 이후, A 대립유전자에 대해서만 양성인 웰, G 대립유전자에 대해서만 양성인 웰, 또는 두 대립유전자 모두에 대해서 양성인 웰의 수를 계수한다. 참조 대립유전자는 보다 적은 수의 양성 웰을 갖는 대립유전자로서 정의된다. 푸아송 확률 밀도 함수에 따라, 다른 대립유전자가 양성인 지에 관계없이, 참조 대립유전자에 대해 음성인 웰의 총 수를 사용하여 m_c 값을 계산한다. 본 발명자의 가설 예의 데이타는 다음과 같다:

[0165] 96-웰 반응에서, 20개의 웰은 A 대립유전자에 대해서만 양성이고, 24개의 웰은 G 대립유전자에 대해서만 양성이며, 33개의 웰은 두 대립유전자 모두에 대해 양성이다. A 대립유전자가 참조 대립유전자로 간주되는데, A-양성 웰이 G-양성 웰에 비해 적기 때문이다. 참조 대립유전자에 대해 음성인 웰의 수는 96 - 20 - 33 = 43이다. 따라서, 푸아송 방정식을 사용하여 다음과 같이 m_c을 계산할 수 있다: -ln(43/96) = 0.80. 이 경우의 실험적으로 결정된 P_1은 24/(20 + 24) = 0.55이다.

[0166] 표 600에 따르면, m_c = 0.8에서의 삼염색체 21 샘플의 예상 P_1은 0.76이다. 따라서, 이 경우 Θ_1은 0.76이다. Θ_1 = 0.76을 기준으로 한 SPRT 곡선을 사용하여 실험적으로 유도된 P_1 (이 경우 0.55)을 해석할 것이다. P_1 = 0.55를 관련된 SPRT 곡선 상에 표시하는 경우, 데이타 지점은 하부 곡선 아래에 있다. 따라서, 본 경우는 정배수체로 분류되며, 도 3을 참조한다.

[0168] 3. 기존 방법과의 비교

[0169] 도 9a는 96-웰 디지탈 RNA-SNP 분석에서의 정배수체 및 삼염색체 21 경우를 분류하기 위한 신규 및 기존 SPRT 알고리즘의 효과의 비교 표 900을 나타낸다. 도 9b는 384-웰 디지탈 RNA-SNP 분석에서 정배수체 및 삼염색체 21 경우를 분류하기 위한 신규 및 기존 SPRT 알고리즘의 효과의 비교의 표 950이다. 신규 알고리즘은 디지탈 PCR 데이타로부터 유도된 m_0에 특이적인 SPRT 곡선의 선택을 의미한다. 기존 알고리즘은 모든 디지탈 PCR 실행에 대한 고정된 세트의 SPRT 곡선의 사용을 의미한다. 분류의 정확도에 대한 분별값의 부정확한 계산의 효과는 표 900에 나타낸 시뮬레이션 분석에 의해 밝혀진다.

[0170] 기존 연구에서 고정된 세트의 SPRT 곡선의 사용과 비교하여, 분류불가능한 데이타의 비율은 표 900 및 950에서 나타난 것과 같이 본 발명의 방법에서 훨씬 적다. 예를 들어, 본 발명의 방법을 사용하여, m_0 = 0.5에서, 14% 및 0%의 삼염색체 21 샘플이 각각 96-웰 및 384-웰 디지탈 RNA-SNP 분석에 대하여 분류불가능할 것이나, 고정된 곡선의 사용에 대해서는 각각 62% 및 10%가 분류불가능할 것이다 (표 900). 따라서, 본 발명의 방법은 보다 적은 수의 정보제공 웰로 질환 분류를 가능하게 한다.

[0171] 표 900에서 나타낸 것과 같이, 신규 알고리즘은 0.1 대지 2.0의 모든 m_0 값에 대한 대립유전자 비율이 벗어나지 나 벗어나지 않은 샘플의 분류에서 보다 정확하다. 예를 들어, m_0이 1.0과 같고, 96-웰 디지탈 RNA-SNP 실행을 수행하는 경우, 신규 알고리즘은 각각 88% 및 82%의, 대립유전자 비율이 벗어난 샘플 및 벗어나지 않은 샘플을 정확하게 분류하며, 반면 기존 알고리즘을 사용하여 대립유전자 비율이 벗어난 샘플 및 벗어나지 않은 샘플의 정확한 분류의 백분율은 단지 각각 19% 및 36%이다.

[0172] 신규 알고리즘을 사용하여, 대립유전자 비율이 벗어난 샘플과 벗어나지 않은 샘플의 분리는 m_0에 따라 증가할 것이다. 결과적으로, 분류 정확도는 m_0에 따라 증가할 것이다. m_0이 2.0 초과로 증가하는 경우 분류 정확도에 대한 2개 샘플 그룹의 분리에서의 증가의 효과는 감소될 것인데, 이는 정보제공 웰의 백분율에서의 감소 때문이다. 반면, 기존 알고리즘의 사용에서, m_0이 증가하는 경우 분류 정확도는 유의하게 감소하는데, 이는 예상 P_0 값과 그의 참값의 편차 증가 때문이다.

[0173] 본 발명자의 실험 및 시뮬레이션 데이타는 디지탈 RNA-SNP가 삼염색체 21 검출에 대해 효과적이며 정확한 방법임을 나타낸다. 모체 혈장에서의 *PLAC1* mRNA가 태아로부터 순수하게 유래되기 때문에, 시험된 12개 모체 혈장 샘플 중 12개에 대하여, 오직 1회의 384-웰 디지탈 PCR 실행만이 정확한 분류에 필요하였다 (도 13b의 표 1350). 이에 따라, 상기 기술한 실시간 디지탈 PCR-기반 방법은 RNA-SNP 분석에 대한 형광 분광법-기반 방법의 또다른 형태를 제공한다 (상기 문헌 [Lo, YMD, et al. 2007 Nat Med]). 태반-특이적 mRNA 전사체와는 별개로, 본 발명자들은 또한 모체 혈장에서의 다른 유형의 태아-특이적 핵산 종을 태아 염색체 이수성의 디지탈 PCR-기반 검출에 사용할 수 있다는 것을 구상한다. 한 예는 태아 후성 마커이며 (문헌 [Chim, SSC et al. (2005) Proc Natl Acad Sci USA 102, 14753-14758]; [Chan, KCA et al. (2006) Clin Chem 52, 2211-2218]), 이는 최근에 후성 대립유전자 비율 (EAR) 방법을 사용하여 삼염색체 18의 비-침습적 출생 전 검출에 사용되었다 (문헌 [Tong, YK et al. (2006) Clin Chem 52, 2194-2202]). 따라서, 본 발명자들은 디지탈 EAR이 가능한 분석 기법일 것으로 예상한다.

[0175] V. 증가 및 다중 마커 및 PCR 반응법

[0176] 상기 기재된 것과 같이, 태아 DNA만이 임신 11주 내지 17주 사이에 대략 3%의 평균 분획 농도로 모체 혈장 DNA의 소수 분획을 구성하는 경우, 모체 혈장으로부터 추출된 DNA에 본 발명의 실시양태를 적용하는 것이 복잡해질 수 있다. 그러나, 본원에서 나타낸 것과 같이, 디지탈 PCR는 심지어 이수체 DNA가 소수 집단으로서 존재하는 경우에도 이수성 검출을 가능하게 한다. 임신 초기 동안 존재할 수 있는, 태아 DNA 분획 농도의 감소에 대하여, 보다 많은 수의 정보제공 계수가 디지탈 PCR에 대해 요구된다. 도 12의 표 1200에서 요약한 것과 같이, 본 작업의 의미는 본 발명자들이 한 세트의 기준 파라미터, 예를 들어 요구되는 태아 DNA 분획 및 총 주형 분자를 제공하였다는 것이며, 이들에 대해 진단 검정이 수립될 수 있다. 본 발명자의 견해에서, 25%의 태아 DNA 분획 농도에 대한 7680개의 총 수의 반응은 목적 예측적인 세트의 기준 파라미터이다. 이러한 파라미터는 표 1200에서 나타낸 것과 같이 정배수체 및 삼염색체 21 샘플이 정확하게 97%로 분류가능하게 한다.

[0177] 모체 혈장의 단위 부피 당 존재하는 혈장 DNA 분자의 수는 제한된다 (문헌 [Lo, YMD, et al. 1998 Am J Hum Genet 62, 768-7758]). 예를 들어, 임신 초기에서, 태아 및 모체 모두로부터의 기여에 대하여, 상업적인 유전

자화, β-글로빈 유전자의 중앙 모체 혈장 농도는 980 카피/mL인 것으로 나타났다 (문헌 [Lo, YMD, et al., 1998 Am J Hum Genet 62, 768-7758]). 7,680개의 분자를 포획하기 위해, 대략 8 mL의 모체 혈장으로부터 추출된 DNA가 필요할 것이다. 약 15 mL의 모체 혈액으로부터 얻을 수 있는 상기 혈장 부피는 일반적인 실시의 한도이다. 그러나, 본 발명자들은 다중 세트의 chr21 및 참조 염색체 표적을 디지탈 RCD 분석을 위해 함할 수 있다는 것을 구상한다. 5쌍의 chr21 및 참조 염색체 표적에 대하여, 단지 1.6 mL의 모체 혈장이 분석에 필요한 주형 분자의 수를 제공하기 위해 필요할 것이다. 멀티플렉스 단일 분자 PCR을 수행할 수 있다. 이러한 멀티플렉스 단일 분자 분석의 견고성(robustness)은 단일 분자 일배체형화(haplotyping)에 대해 이미 입증되었다 (문헌 [Ding, C. and Cantor, CR. 2003 Proc Natl Acad Sci USA 100, 7449-7453]).

[0178] 별법으로, 25%의 태아 DNA 분획 농도를 달성하기 위해, 방법은 모체 혈장에서 태아 DNA의 선택적 풍부화 (문헌 [Li, Y. et al. 2004 Clin Chem 50, 1002-1011]) 또는 모체 DNA 배경의 억제 (문헌 [Dhallan, R et al. 2004 JAMA 291, 1114-1119]) 또는 둘 모두를 가능하게 할 수 있다. 태아 DNA 풍부화 및 모체 DNA 억제를 위한 물리적 방법과는 별개로, 특정 DNA 메틸화 패턴을 나타내는 태아 DNA 분자의 표적화와 같은 분자 풍부화 전략을 사용하는 것이 또한 가능할 것이다 (문헌 [Chim, SSC et al. 2005 Proc Natl Acad Sci USA 102,14753-14758]; [Chan, KCA et al. 2006 Clin Chem 52, 2211-2218]; [Chiu, RWK et al. 2007 Am J Pathol 170, 941-950.]).

[0179] 추가적으로, 디지탈 PCR 수행에 대한 현행 연구에 사용되는 것과 같은 디지탈 실시간 PCR 분석의 수동 설정을 위한 다수의 또다른 방법이 현재 존재한다. 이러한 다른 방법에는 미세유체 디지탈 PCR 칩 (문헌 [Warren, L et al., 2006 Proc Natl Acad Sci USA 103, 17807-17812]; [Ottesen, EA et al. 2006 Science 314, 1464-1467]), 에멀젼 PCR (문헌 [Dressman, D et al. 2003 Proc Natl Acad Sci USA 100, 8817-8822]), 및 예를 들어 로슈(Roche) 454 플랫폼, 일루미나 솔렉사(Illumina Solexa) 플랫폼, 및 어플라이드 바이오시스템즈(Applied Biosystems)의 솔리드(SOLiD)™ 시스템을 사용하는 대규모 병렬 게놈 서열분석 (문헌 [Margulies, M. et al. 2005 Nature 437, 376-380]) 등이 있다. 추가에 대하여, 본 발명자의 방법은 또한 단일 DNA 분자에 대한 대규모 병렬 서열분석 방법에 적용가능하며, 이는 증폭 단계, 예컨대 헬리코스 트루(Helicos True) 단일 분자 DNA 서열분석 기술 (문헌 [Harris TD et al., 2008 Science, 320, 106-109]), 퍼시픽 바이오사이언시즈(Pacific Biosciences)의 단일 분자, 실시간 (SMRT™) 기술, 및 나노공극(nanopore) 서열분석 (문헌 [Soni GV and Meller A. 2007 Clin Chem 53, 1996-2001])을 요구하지 않는다. 이들 방법을 사용하여, 디지탈 RNA-SNP 및 디지탈 RCD를 다수의 샘플에 대하여 빠르게 수행함으로써, 비-침습적 출생 전 진단에 대해 본원에서 제안된 방법의 임상적 실행가능성을 향상시킬 수 있다.

도면의 간단한 설명

[0181] 도 1은 디지탈 PCR 실험을 설명하는 초롬도이다.

도 2a는 본 발명의 실시양태에 따른 디지탈 RNA-SNP 및 RCD 방법을 예시한다.

도 2b는 양에서 자주 검출되는 염색체 이상의 예에 대한 표를 도시한다.

도 3은 본 발명의 실시양태에 따른 다운 증후군을 결정하는데 사용된 SPRT 곡선을 갖는 그래프를 예시한다.

도 4는 본 발명의 실시양태에 따른 태아 세포의 백분율을 사용하여 질환 상태를 결정하는 방법을 도시한다.

도 5는 본 발명의 실시양태에 따른 평균 농도를 사용하여 질환 상태를 결정하는 방법을 도시한다.

도 6은 본 발명의 실시양태에 따른 웰 당 평균 참조 주형 농도 (m_c)로서 표현된 주형 농도 범위에 대해 삼염색체 21 샘플의 예상되는 디지탈 RNA-SNP 대립유전자 비율 및 P_A을 표로 만든 표를 도시한다.

도 7은 본 발명의 실시양태에 따른 웰 당 평균 참조 주형 농도 (m_c)로서 표현된 주형 농도 범위에서 삼염색체 21 샘플 중 분획 태아 DNA 농도 10%, 25%, 50% 및 100%에 대해 예상되는 P_A을 표로 만든 표를 도시한다.

도 8은 본 발명의 실시양태에 따른 디지탈 RNA-SNP 분석에서 m_c 값 0.1, 0.5 및 1.0에 대해 SPRT 곡선의 차이 정도를 예시하는 플롯을 도시한다.

도 9a는 본 발명의 실시양태에 따른 96-웰 디지탈 RNA-SNP 분석에서 정배수체 및 삼염색체 21 사례를 분류하기 위한 신규/기존 SPRT 알고리즘 효과의 비교에 대한 표를 도시한다.

도 9b는 본 발명의 실시양태에 따른 384-웰 디지탈 RNA-SNP 분석에서 정배수체 및 삼염색체 21 사례를 분류하기

위한 신규/기존 SPRT 알고리즘 효과의 비교에 대한 표를 도시한다.

도 10은 본 발명의 실시양태에 따른 소정의 염색체 서열에 대해 정배수체 또는 이수체로서 정확하게 및 부정확하게 분류된 태아 및 분류가능하지 않은 태아의 백분율을 도시하는 표이다.

도 11은 본 발명의 실시양태에 따른 순수 (100%) 태아 DNA 샘플에 대한 디지털 RCD 분석용 컴퓨터 시뮬레이션을 도시하는 표 1100이다.

도 12는 본 발명의 실시양태에 따른 상이한 분획 농도의 태아 DNA를 갖는 정배수체 또는 삼염색체 21 태아로부터의 샘플을 분류하기 위해 α_c=0.5에서 디지털 RCD 분석의 정확성을 컴퓨터 시뮬레이션한 결과를 도시하는 표 1200이다.

도 13a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 태반 조직에서의 디지털 RNA-SNP 분석에 대한 표 1300을 도시한다.

도 13b는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신으로부터의 모체 혈장의 디지털 RNA-SNP 분석에 대한 표 1350을 도시한다.

도 14a 내지 14c는 본 발명의 실시양태에 따른 RCD 분석으로부터 생성된 분별값 곡선을 예시하는 플롯을 도시한다.

도 15a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 태반 조직에서의 디지털 RNA-SNP 분석에 대한 표를 도시한다.

도 15b는 본 발명의 실시양태에 따른 한 모체 혈장 샘플로부터의 12개 반응 웰들의 디지털 RNA-SNP 데이터에 대한 표를 도시한다.

도 15c는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신으로부터의 모체 혈장의 디지털 RNA-SNP 분석에 대한 표를 도시한다.

도 16a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 18 태반의 디지털 RNA-SNP 분석에 대한 표를 도시한다.

도 16b는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 18 태반에 대한 디지털 RNA-SNP 데이터의 SPRT 해석을 도시한다.

도 17은 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신의 30% 태반/모체 혈액 세포 DNA 혼합물의 디지털 RCD 분석에 대한 표를 도시한다.

도 18은 본 발명의 실시양태에 따른 정확한 분류에 대한 판정 경계값을 예시하는 SPRT 곡선을 도시한다.

도 19는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신으로부터의 양수 샘플의 디지털 RCD 분석에 대한 표를 도시한다.

도 20은 본 발명의 실시양태에 따른 정배수체 및 삼염색체 18 임신 (E = 정배수체; T18 = 삼염색체 18)으로부터의 태반 DNA 샘플의 디지털 RCD 분석에 대한 표를 도시한다.

도 21은 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 (E = 정배수체; T21 = 삼염색체 21; U = 미분류)의 30% 태반/모체 혈액 세포 DNA 혼합물의 멀티플렉스 디지털 RCD 분석에 대한 표를 도시한다.

도 22a 및 22b는 본 발명의 실시양태에 따른 50% 정배수체 또는 삼염색체 21 태반 게놈 DNA/50% 모체 연막(buffy coat) DNA 혼합물의 멀티플렉스 디지털 RCD 분석에 대한 표를 도시한다. 미분류(Unclass)는 분류가능하지 않음을 나타내고, T21은 삼염색체 21을 나타낸다.

도 23은 남성 및 여성 파트너 둘 다가 동일한 돌연변이를 전달하는 시나리오를 도시한다.

도 24a는 본 발명의 실시양태에 따른 여성/남성 및 남성/남성 DNA 혼합물의 디지털 RMD 분석에 대한 표를 도시한다.

도 24b는 본 발명의 실시양태에 따른 25% 여성 및 75% 남성 DNA를 갖는 혼합물의 디지털 RMD 분석에 대한 표를 도시한다.

도 25는 본 발명의 실시양태에 따른 HbE 돌연변이에 대해 모체 혈장 샘플을 모방하는 15% 내지 50% DNA 혼합

물의 디지탈 RBD 분석에 대한 표를 도시한다.

도 26a는 본 발명의 실시양태에 따른 CD41/42 돌연변이에 대해 모체 혈장 샘플을 모방하는 5% 내지 50% DNA 혼합물의 디지탈 RBD 분석에 대한 표를 도시한다.

도 26b는 본 발명의 실시양태에 따른 CD41/42 돌연변이에 대해 모체 혈장 샘플을 모방하는 20% DNA 혼합물의 디지탈 RBD 분석에 대한 표를 도시한다.

도 27은 본 발명의 실시양태에 따른 시스템 및 방법으로 이용가능한 예시적 컴퓨터 장치의 물리 다이아그램을 도시한다.

발명을 실시하기 위한 구체적인 내용

[0182] 실시예

[0183] 하기 실시예는 청구된 발명을 예시하되 제한하지 않으려고 한다.

[0185] I. 컴퓨터 시뮬레이션

[0186] SPRT 방법을 사용하여 삼염색체 21 진단의 정확도를 평가하기 위해 컴퓨터 시뮬레이션을 수행하였다. 마이크로소프트 엑셀 2003 소프트웨어 (마이크로소프트 코포레이션(Microsoft Corp.; 미국)) 및 윈도우용 SAS 9.1 소프트웨어 (에스에이에스 인스티튜트 인코.(SAS Institute Inc., 미국 노스 캐롤라이나주))를 사용하여 컴퓨터 시뮬레이션을 수행하였다. 디지탈 PCR의 수행은 참조 주형 농도 (m_c), 정보제공 계수의 수, 및 대립유전자 또는 염색체 불균형의 추정 정도 (P_c) 사이의 상호작용이다. 이들 변수 각각의 범위에 대하여 별도의 시뮬레이션을 수행하였다. 디지탈 RNA-SNP 및 디지탈 RCD에 대한 SPRT 곡선의 결정 경계가 상이하기 때문에, 2개 시스템에 대한 시뮬레이션 분석을 별도로 수행하였다.

[0187] 시뮬레이션된 각각의 디지탈 PCR 조건 (즉, m_c, 태아 DNA 분획 농도, 총 웰 수)에 대하여, 2회의 시뮬레이션을 수행하였다. 제1회는 시험된 샘플을 정배수체 태아를 갖는 임신 여성으로부터 얻은 시나리오를 시뮬레이션하였다. 제2회는 시험된 샘플을 삼염색체 21 태아를 갖는 임신 여성으로부터 얻은 경우의 시나리오를 시뮬레이션하였다. 각각의 시뮬레이션에서, 5000명의 태아의 시험을 시뮬레이션하였다.

[0189] A. RNA-SNP

[0190] 디지탈 RNA-SNP에 대하여, m_c = 0.1 내지 m_c = 2.0으로 384-웰 실험의 시뮬레이션을 수행하였다. 각각의 m_c 값에서, 본 발명자들은 5000명의 정배수체 및 5000명의 삼염색체 21 태아를 시험하는 시나리오를 시뮬레이션하였다. 주어진 m_c에 적절한 SPRT 곡선을 사용하여 10,000명의 태아를 분류하였다. 도 10은 본 방법의 실시양태에 따른, 정배수체 또는 이수체로서 정확하게 및 부정확하게 분류된 태아, 및 주어진 정보제공 계수에 대해 분류가 능하지 않은 태아의 백분율을 나타내는 표 1000이다. 정배수체 및 이수체 경우의 진단에 대한 정확도는 0.5 내지 2.0 사이의 m_c에 대해 모두 100%이다. m_c = 0.1인 경우, 57% 및 88%의 정배수체 및 삼염색체 21 태아만이 384-웰의 분석 이후 정확하게 분류될 수 있었다.

[0191] 하기 단계에 기재된 것과 같이 시뮬레이션 데이타를 생성하였다.

[0192] 단계 1에서, 각 웰에 대하여, SAS 프로그램의 무작위 (푸아송) 함수 (www.sas.com/technologies/analytics/statistics/index.html)를 사용하여 2개 난수를 생성하여 각각 A 및 G 대립유전자를 나타내었다. 무작위 (푸아송) 함수는 0으로부터 출발하는 양의 정수 (즉, 0, 1, 2, 3,...)를 생성할 것이며, 생성되는 각 정수의 확률은 웰 당 대립유전자의 평균 농도를 나타내는 주어진 평균값에 대한 푸아송 확률 밀도 함수에 따른 상기 수의 확률에 의존하였다. A 대립유전자를 나타내는 난수가 0보다 큰 경우, 즉 A 대립유전자의 1개 이상의 분자를 함유한 경우, 웰은 A 대립유전자에 대해 양성으로서 간주되었다. 유사하게, G 대립유전자를 나타내는 난수가 0보다 큰 경우, 웰은 G 대립유전자에 대해 양성으로서 간주되었다.

[0193] 정배수체 태아를 갖는 임신 여성의 시나리오를 시뮬레이션하기 위해, A 대립유전자 및 G 대립유전자에 대한 난수를 생성하는데 동일한 평균 값을 사용하였다. 예를 들어, m_c = 0.5에서의 디지탈 RNA-SNP 분석을 시뮬레이션하는 분석에서, A 대립유전자 또는 G 대립유전자 중 하나에 대한 평균 값을 웰 당 0.5 분자의 대립유전자 중 하나에 대한 평균 농도를 의미하는 0.5와 동일하게 설정하였다. 푸아송 방정식을 사용하여, 0.5의 평균

농도에서, A 또는 G 대립유전자에 대해 양성인 웰의 비율은 동일할 것이며, 0.3935였다 (표 600 참조).

[0194] m_c = 0.5에서 삼염색체 21 태아를 갖는 임신 여성의 디지탈 RSA-SNP 분석을 시뮬레이션한 경우, 웰 당 과다표출된 대립유전자의 평균 농도는 참조 대립유전자의 평균 농도의 2배, 즉 1이 되는 것으로 예상될 것이다. 이 상황에서, 과다표출된 대립유전자에 대해 양성인 웰의 확률은 0.6321이었다 (표 600 참조).

[0195] 디지탈 PCR 웰에 대한 난수의 생성 이후, 웰을 하기 상태 중 하나로서 분류할 수 있었다:

[0196] a. A 및 G 대립유전자 모두에 대해 음성

[0197] b. A 및 G 대립유전자 모두에 대해 양성

[0198] c. A 대립유전자에 대해 양성이되, G 대립유전자에 대해 음성

[0199] d. G 대립유전자에 대해 양성이되, A 대립유전자에 대해 음성.

[0200] 단계 2에서, 원하는 수의 웰 (현재 시뮬레이션에 대하여 384-웰)을 생성할 때까지 단계 1을 반복하였다. A 대립유전자에 대해서만 양성인 웰의 수 및 G 대립유전자에 대해서만 양성인 웰의 수를 계수하였다. 덜 양성인 웰을 갖는 대립유전자는 참조 대립유전자로서 간주되며, 보다 양성인 웰을 갖는 대립유전자는 잠재적으로 과다표출된 대립유전자로서 간주되었다. 정보제공 웰의 수는 대립유전자 중 하나에 대해 양성이되, 모두에 대해서는 아닌 웰의 총 수였다. 이후, 잠재적으로 과다표출된 대립유전자를 함유하는 정보제공 웰의 비율 (P_v)을 계산하였다. 귀무가설 또는 대립가설을 수용하기 위한 관련 SPRT 곡선에 대한 상한 및 하한을 본 발명의 실시양태에 따라 계산하였다.

[0201] 단계 3에서, 정배수체 또는 삼염색체 21 태아를 갖는 임신 여성의 각각의 2개 시나리오에 대해 3000의 시뮬레이션을 수행하였다. 각각의 시뮬레이션은 임신 여성으로부터 얻어진 독립적인 생물학적 샘플로서 간주될 수 있었다. 표 1000에서, 정배수체 경우의 정확한 분류는 귀무가설이 수용되는 정배수체 경우를 의미하며, 정배수체 경우의 부정확한 분류는 대립가설이 수용되는 정배수체 경우를 의미한다. 유사하게, 대립가설이 수용되는 삼염색체 21 경우는 정확하게 분류된 것으로 간주되며, 귀무가설이 수용되는 삼염색체 21 경우는 부정확하게 분류된 것으로 간주된다. 두 군 모두에 대하여, 미타-영사된 웰의 총 수를 시뮬레이션한 이후 귀무가설 또는 대립가설 모두가 수용되지 않는 경우에, 미분류된 것으로 간주되었다.

[0202] 단계 4에서, 단계 1 내지 3을 0.1 내지 2.0 범위의 m_c에 대해 0.1 증가량으로 수행하였다.

[0204] B. RCD

[0205] 도 11은 본 발명의 실시양태에 따라 0.1 내지 2.0 범위의 m_c에서 순수한 (100%) 태아 DNA 샘플에 대한 디지탈 RCD 분석을 컴퓨터 시뮬레이션을 나타내는 표 1100이다. 태아 DNA 분획 농도가 낮아짐에 따라, 염색체 21 과다표출 정도는 감소되며, 이에 따라 정확한 질환 분류를 위해 보다 많은 수의 정보제공 웰이 요구된다. 따라서, m_c = 0.5에서 384개 내지 7680개 웰 범위의 총 웰 수에 대하여 50%, 25% 및 10%의 태아 DNA 농도에 대해 추가로 시뮬레이션을 수행하였다.

[0206] 도 12는 본 발명의 실시양태에 따른 상이한 분획 농도의 태아 DNA를 갖는 정배수체 또는 삼염색체 21 태아로부터의 샘플의 분류에 대하여 m_c = 0.5에서 디지탈 RCD 분석의 정확도의 컴퓨터 시뮬레이션 결과를 나타내는 표 1200이다. 디지탈 RCD의 수행은 보다 높은 태아 DNA 분획 농도를 갖는 경우에 대해 보다 양호하다. 25% 태아 DNA 농도 및 7680개 총 수의 PCR 분석에서, 97%의 정배수체 및 이수체 경우 모두가 부정확한 분류 없이 분류가 능할 것이다. 나머지 3%의 경우는 분류가 달성될 수 있을 때까지 추가 분석을 요구한다.

[0207] 디지탈 RCD 분석을 시뮬레이션하기 위한 절차는 디지탈 RSA-SNP 분석에 대해 기재된 것과 유사하였다. 시뮬레이션을 위한 단계를 하기 기재한다.

[0208] 단계 1에서, 푸아송 확률 밀도 함수 하의 2개의 난수를 생성하여 참조 유전자좌, 염색체 1 및 염색체 21 유전자좌를 나타내었다. 정배수체 태아를 갖는 대상체에 대하여, 염색체 1 및 염색체 21 유전자좌 모두의 평균 농도는 동일하였다. 상기 시뮬레이션 분석에서, 웰 당 각 유전자좌에 대한 0.5의 평균 주형 농도를 사용하였다. 삼염색체 21 태아를 갖는 대상체에 대하여, 상기 시뮬레이션에서의 m_c은 0.5였으나, 웰 당 염색체 21 유전자좌의 평균 농도는 표 700에서 나타낸 것과 같이 시험된 샘플에서의 태아 DNA 분획 농도에 의존할 것이다. 웰에 대한 참조 및/또는 염색체 21 유전자좌의 분포를 웰 당 유전자좌의 적절한 평균 농도를 갖는 푸아송 확률 밀도

함수에 따라 생성된 각각의 유전자좌를 나타내는 난수에 의해 측정하였다.

[0209] 단계 2에서, 원하는 수의 웰, 예를 들어 384-웰 플레이트 실험에 대하여 384개 웰이 생성될 때까지 단계 1을 반복하였다. 염색체 1 단독 및 염색체 21 단독에 대해서만 양성인 웰의 수를 계수하였다. 정보제공 웰의 수는 염색체 중 하나에 대해 양성이며, 모두에 대해서는 아닌 웰의 총 수였다. 이후, 염색체 21에 대해 양성인 정보제공 웰의 비율 (P_i)을 계산하였다. 귀무가설 또는 대립가설을 수용하기 위한 관련 SPRT 곡선의 상한 및 하한을 SPRT 분석에 대해 앞선 섹션에서 기재된 것과 같이 계산하였다.

[0210] 단계 3에서, 정배수체 또는 삼염색체 21 태아를 갖는 임신 여성의 2개 시나리오 각각에 대해 5000회의 시뮬레이션을 수행하였다. 각각의 시뮬레이션은 임신 여성으로부터 얻어진 독립적인 생물학적 샘플로서 간주될 수 있었다. 표 1100에서, 정배수체 경우의 정확한 분류는 귀무가설이 수용되는 정배수체 경우를 의미하며, 정배수체 경우의 부정확한 분류는 대립가설이 수용되는 정배수체 경우를 의미한다. 유사하게, 대립가설이 수용되는 삼염색체 21 경우는 정확하게 분류된 것으로 간주되며, 귀무가설이 수용되는 삼염색체 21 경우는 부정확하게 분류된 것으로 간주되었다. 두 군 모두에 대하여, 미리-명시된 웰의 총 수를 시뮬레이션한 이후 귀무가설 또는 대립가설 모두가 수용되지 않는 경우에, 미분류된 것으로 간주되었다.

[0211] 단계 4에서, 단계 1 내지 3을 384개 내지 7680개 범위의 총 웰 수에서 10% 25%, 50% 및 100% 태아 DNA를 갖는 샘플에 대해 반복하였다.

[0213] 11. 삼염색체 21 검출의 검증

[0214] A. PLAC4에 대한 RNA-SNP

[0215] 염색체 21 상의 PLAC4 유전자에 대한 rs8130833 SNP를 사용하여 디지털 RNA-SNP의 실제적 실행가능성을 입증하였다 (문헌 [Lo, YMD et al. 2007 Nat Med 13, 218-223]). 2개 정배수체 및 2개 삼염색체 21 이형접합 태반으로부터의 태반 DNA 및 RNA 샘플을 분석하였다. 디지털 RNA-SNP 프로토콜을 사용하되, 역전사 단계를 생략하였으며, 이에 따라 필수적으로 경사를 디지털 DNA-SNP 분석으로 전환하여, 태반 DNA 샘플을 분석하였다. 정확한 경우 분류의 가능성과 정보제공 웰의 비율 사이에 균형을 맞추기 위해, 본 발명자들은 웰 당 입의의 유형의 하나의 대립유전자를 목적으로 하는 샘플을 희석하고, 96-웰 디지털 PCR 분석에 의해 확인하였다. 이후, 384-웰 디지털 RNA-SNP 실험을 하였다. P_i 및 m_i을 계산하고, 상기 m_i 값에 대한 SPRT 곡선을 데이타 해석을 위해 사용하였다.

[0216] 도 13a는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신의 태반 조직에서의 디지털 RNA-SNP 분석의 표 1300을 나타낸다. 실험적으로 얻어진 P_i이 분류불가능한 영역 아래에 있는 경우 "정배수체"로 지정하였고, 실험적으로 얻어진 P_i이 분류불가능한 영역 위에 있는 경우 삼염색체 21을 나타내는 "T21"로 지정하였다. T21, 삼염색체 21. 1회 384-웰 실험으로 DNA 및 RNA 샘플 모두를 사용하여 각각의 상기 경우를 정확하게 분류하였다.

[0217] 추가적으로, 본 발명자들은 정배수체 태아를 갖는 9명의 여성 및 삼염색체 21 태아를 갖는 4명의 여성으로부터의 혈장 RNA 샘플을 시험하였다. 도 13b는 본 발명의 실시양태에 따른 정배수체 및 삼염색체 21 임신 여성으로부터의 모체 혈장의 디지털 RNA-SNP 분석의 표 1350을 나타낸다. 모든 경우를 정확하게 분류하였다. 한 삼염색체 21 경우 (R227QP)에 대한 초기 결과는 1회 384-웰 실험 이후 SPRT 곡선 사이의 분류불가능한 영역 내에 있었다. 따라서, 추가의 384-웰 실행을 수행하였다. 총 768개 웰로부터의 응집된 데이타로부터 신규 m_i 및 P_i 값을 계산하였으며, 상기 m_i 값을 기준으로 선택된 신규 세트의 SPRT 곡선을 사용하여 분류를 수행하였다. 이후, 상기 경우를 이수체로서 정확하게 스코어링하였다.

[0218] 본 발명자의 실험 및 시뮬레이션 데이타는 디지털 RNA-SNP가 삼염색체 21 검출에 대한 효과적이며 정확한 방법임을 나타낸다. 모체 혈장 중 *PLAC4* mRNA가 태아로부터 순수하게 유래되기 때문에, 시험된 13개 모체 혈장 샘플 중 12개에 대하여, 오직 1회의 384-웰 디지털 PCR 실행만이 정확한 분류를 위해 요구되었다. 이에 따라, 상기 균질한 실시간 디지털 PCR-기반 방법은 RNA-SNP 분석을 위한 질량 분광법-기반 방법과는 또다른 방법을 제공한다. 태반-특이적 mRNA 전사체와는 별개로, 본 발명자들은 또한 모체 혈장에서의 다른 유형의 태아-특이적 핵산 중의 태아 염색체 이수성의 디지털 PCR-기반 검출에 대해 사용될 수 있다는 것을 구상한다. 한 예는 최근에 후성 대립유전자 비 (EAR) 방법을 사용하여 삼염색체 18의 비-침습적 출생 전 검출에 사용된 태아 후성 마커이다 (문헌 [Tong YK et al. 2006 Clin Chem, 52, 2194-2202]). 따라서, 본 발명자들은 디지털 EAR이 가능한 분

서 기술될 것이라고 예상한다.

[0220] B. RCD

[0221] 또한, 염색체 21 및 1에서의 파라로그 서열을 표적화하는 PCR 분석을 사용하여 삼염색체 21 검출에 대한 디지탈 RCD의 실제적 실행가능성을 연구하였다. 파라로그 유전자좌를 본원에서 예로서 사용하였다. 염색체 21 및 임의의 다른 참조 염색체에 대한 비-파라로그 서열 또한 RCD에 대해 사용할 수 있다. 2개 정배수체 및 2개 삼염색체 21 태반으로부터의 태반 DNA 샘플을 웰 당 염색체 중 하나로부터 대략 1개 표적 주형으로 희석시켰으며, 96-웰 디지탈 PCR 분석에 의해 확인하였다. 각각의 확인된 샘플을 384-웰 디지탈 RCD 실험에 의해 분석하였으며, P, 및 m, 값을 계산하였다. 디지탈 RCD에 대하여, 염색체 1 파라로그는 참조 주형이었다. 상기 m, 값을 사용하여 데이타 해석을 위한 상응하는 세트의 SPRT 파선을 선택하였다. 모든 태반 샘플을 도 14a에서 나타낸 것과 같이 정확히 분류하였다.

[0222] 모체 혈장에서의 태아 DNA의 시나리오와 같은 디지탈 RCD 방법을 사용하여 과량의 정배수체 DNA와 혼합된 삼염색체 21 DNA를 검출할 수 있는지를 입증하기 위해, 정배수체 모체 혈액 세포 DNA의 배경 중 50% 및 25%의 삼염색체 21 태반 DNA를 함유하는 혼합물을 분석하였다. 10개의 삼염색체 21 및 10개의 정배수체 공우로부터의 태반 DNA를 각각 동일한 양의 정배수체 모체 혈액 세포 DNA와 혼합하였으며, 이에 따라 20개의 50% DNA 혼합물을 제조하였다. 도 14b는 본 발명의 실시양태에 따른 50% 태아 DNA 혼합물의 RCD 분석에 대한 SPRT 해석을 예시하는 플롯 (1440)을 나타낸다. 유사하게는, 5개의 삼염색체 21 및 5개의 정배수체 공우로부터의 태반 DNA를 각각 3배 과량의 정배수체 모체 혈액 세포 DNA와 혼합하였으며, 이에 따라 10개의 25% DNA 혼합물을 제조하였다. 도 14c는 25% 태아 DNA 혼합물의 RCD 분석에 대한 SPRT 해석을 예시하는 플롯 1470을 나타낸다. 도 14b 및 14c에서 나타낸 것과 같이, 모든 정배수체 및 이수체 DNA 혼합물을 정확하게 분류하였다.

[0223] 각각의 샘플은 도 14b 및 14c에 표시된 것과 같이 여러 384-웰 디지탈 PCR 분석 이후 분류가능한 지점에 도달하였다. 50% DNA 혼합물에 대하여, 384-웰 플레이트의 수는 1개 내지 5개의 범위가 요구된다. 25% DNA 혼합물에 대하여, 384-웰 플레이트의 수는 1개 내지 7개의 범위가 요구된다. 매 384회의 디지탈 PCR 분석의 추가와 함께 정확하게 분류되는 경우의 누적 비율은 표 1200에서 제시된 컴퓨터 시뮬레이션에 의해 예상된 것과 같았다.

[0225] III. 디지탈 PCR을 이용한 발명

[0226] A. 디지탈 RNA-SNP

[0227] 모든 RNA 샘플을 먼저 써모스크립트(ThermoScript) 역전사효소 (인비트로젠(Invitrogen))를 사용하여 유전가-특이적 역전사 프라이머에 의해 역전사시켰다. 역전사 프라이머의 서열은 5'-AGTATATAGAACCATGTTTAGGCCAGA-3' (서열 1) (미국 아이오와주 코랄빌 소재의 인테그레이티드 DNA 테크놀러지스(Integrated DNA Technologies))였다. 디지탈 RNA-SNP를 위한 역전사된 RNA (즉, cDNA) 샘플 및 DNA 샘플 (예를 들어, 태반 DNA)의 후속 처리는 본질적으로 동일하였다. 디지탈 PCR 분석 전에, 먼저 DNA 및 cDNA 샘플을 프라이머 5'-CCGCTAGGGTGTCTTTTAAGC-3' (서열 2), 5'-GTGTTGCAATACAAAATGAGTTTCT-3' (서열 3), 및 형광 프로브 5'-(FAM)ATTGGAGCAAATTC(MGBNFQ)-3' (서열 4) (미국 캘리포니아주 포스터 시티 소재의 어플라이드 바이오시스템즈) (여기서, FAM은 6-카르복시플루오레세인이고, MGBNFQ는 마이너 그루브 결합 비-형광 켄처임)로 이루어진, PLAC4에 대한 실시간 PCR 검정을 사용하여 정량화하였다.

[0228] 샘플링은이 특정된 HPLC-정제된 단일 가닥 합성 DNA 올리고뉴클레오티드 (성가포르 소재의 프롤리고(Proligo)의 계열 희석에 의해 보정 곡선을 제작하였다. 서열은 5'-CGCCGCTAGGGTGTCTTTTAAGCTATTGGAGCAAATTCAAATTTGGCTTAAAGAA AAAGAAACTCATTTTGTATTGCAACACCAGGAGTATCCCAAGGGACTCG-3' (서열 5)였다. 반응 부피 25 μL로 2X 테크만 유니버셜 PCR 마스타 믹스(TaqMan Universal PCR Master Mix) (어플라이드 바이오시스템즈)를 사용하여 반응을 설정하였다. 각각의 프라이머 400 nM 및 프로브 80 nM을 각각의 반응에 사용하였다. ABI PRISM 7900HT 서열 검출 시스템 (어플라이드 바이오시스템즈)에서, 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 60℃에서 1분의 45주기를 수행하였다. 그 다음, 웰 당 대략 1개의 주형 분자에서 후속 디지탈 PCR 증폭이 수행될 수 있도록 DNA 또는 cDNA 샘플의 계열 희석에

하수하였다. 이러한 농도에서는, 반응 웰의 대략 37%가 음성 증폭을 나타낼 것으로 예측되며, 이는 96-웰 디지탈 실시간 PCR 분석을 수행함으로써 최초로 확인되었다. 이어서, 한 세트의 비-인트론 분할(spanning) 프라이머: 전방향 프라이머 **5'-TTTGTATTGCAACACCATTTGG-3'** (서열 6) 및 앞서 기재된 유전자-특이적 역전사 프라이머를 사용하여 384-웰 플레이트에서 디지탈 RNA-SNP 분석을 수행하였다.

[0229] PLAC4 서열 상 rs8130833 SNP의 2개의 대립유전자를 각각 표적으로 하는 2개의 대립유전자-특이적 테크만 프로브를 설계하였다. 그들의 서열은 G 및 A 대립유전자에 대하여 각각 **5'- (FAM)TCGTCGTCTAACTTG(MGBNFQ)-3'** (서열 7) 및 5'-(VIC)ATTCGTCATCTAACTTG(MGBNFQ)-3' (서열 8)였다. 반응 부피 5 μL로 2X 테크만 유니버셜 PCR 마스터 믹스를 사용하여 반응을 설정하였다. 각각의 반응은 1X 테크만 유니버셜 PCR 마스터 믹스, 각각의 프라이머 572 nM, 대립유전자-G-특이적 프로브 107 nM 및 대립유전자-A-특이적 프로브 357 nM을 함유하였다. 반응은 ABI PRISM 7900HT 서열 검출 시스템에서 수행하였다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 45주기를 수행하였다. 반응 동안, SDS 2.2.2 소프트웨어 (어플라이드 바이오시스템즈)의 "절대 정량화" 응용 프로그램에 의해 형광 데이타를 수집하였다. 상기 소프트웨어는 기준선 및 임계값을 자동으로 계산하였다. A 또는 G 대립유전자에 대해 양성인 웰의 수를 기록하여 SPRT 분석에 제공하였다.

[0231] B. 디지탈 RCD 분석

[0232] 본 연구에 사용되는 모든 태반 및 모체 혈장 DNA 샘플을 먼저 나노드롭 분광광도계 (미국 델라웨어주 윌밍턴 소재의 나노드롭 테크놀러지(NanoDrop Technology))에 의해 정량화하였다. 6.6 pg/세포의 전환을 사용하여 DNA 농도를 카피/μL로 전환하였다. 웰 당 대략 1개의 주형에 상응하는 DNA의 양을, DNA 샘플을 제열 최적함으로써 결정하고 96-웰 포맷에서 실시간 PCR 검정에 의해 확인하였는데, 여기서 웰의 대략 37%가 음성 증폭을 나타낼 것으로 예측되었다. 확인 플레이트에 대한 PCR 설정은 참조 염색체에 대한 프로브를 추가한 것만 제외하고는 이하에 기재된 것과 동일하였다. 디지탈 RCD 분석에서, 염색체 21 및 1 상의 파라로그 유전자와 (문헌 [Deutsch, S. et al. 2004 J Med Genet 41, 908-915])를 먼저 전방향 프라이머 **5'-GTTGTTCTGCAAAAAACCTTCGA-3'** (서열 9) 및 역방향 프라이머 **5'-CTTGGCCAGAAATACTTCATTACCATAT-3'** (서열 10)에 의해 동시 증폭시켰다. 염색체 21 및 1 파라로그를 표적으로 하도록 2개의 염색체-특이적 테크만 프로브를 설계하였는데, 그들의 서열은 각각 **5'-(FAM)TACCTCCAT AATGAGTAAA(MGBNFQ)-3'** (서열 11) 및 **5'-(VIC)CGTACCTCTGTAATGTGTAA(MGBNFQ)-3'** (서열 12)였다. 각각의 반응은 1X 테크만 유니버셜 PCR 마스터 믹스 (어플라이드 바이오시스템즈), 각각의 프라이머 450 nM 및 각각의 프로브 125 nM을 함유하였다. 총 반응 부피는 5μL/웰이었다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 60℃에서 1분의 50주기를 수행하였다. 모든 실시간 PCR 실험을 ABI PRISM 7900HT 서열 검출 시스템 (어플라이드 바이오시스템즈) 상에서 수행하였고, SDS 2.2.2 소프트웨어 (어플라이드 바이오시스템즈)의 "절대 정량화" 응용 프로그램에 의해 형광 데이타를 수집하였다. 디폴트 기준선 및 매뉴얼 임계값을 사용하였다. 염색체 21 또는 염색체 1에 대해 양성인 웰의 수를 기록하여 SPRT 분석에 제공하였다. SPRT에 의해 결론 분류가 가능해질 때까지 1개 이상의 384-웰 플레이트가 분석될 것이다.

[0234] <u>IV. 미세유체-기반 디지탈 PCR의 사용</u>

[0235] A. 디지탈 RNA-SNP

[0236] 본 실시예는 미세유체-기반 디지탈 PCR을 사용한 디지탈 PCR 분석의 성능을 증명한다. 플루이다임 바이오마크(Fluidigm BioMark)™ 시스템을 사용하는 이 접근법의 변형이 그에 제한되지 않는 예로서 본원에 예시되었다. 상기 시스템은 실행 1회당 9000회 초과의 디지탈 PCR을 수행할 수 있다.

[0237] 정해수재 또는 산염색체 21 태아를 가진 임신 여성으로부터 태반 조직 및 모체 말초 혈액 샘플을 채취하였다. 프라이머 연장에 이은 정량 분광법에 의해 태반 DNA 샘플에서 PLAC4 유전자 상 rs8130833 SNP의 유전자형 결정을 수행하였다. 태반 및 모체 혈장 샘플로부터 RNA를 추출하였다.

[0238] 모든 RNA 샘플을 씨모스크립트 역전사효소 (인비트로젠)를 사용하여 유전자-특이적 역전사 프라이머 (**5'-AGTATATAGAACCATGTTTAGGCCAGA-3'** (서열 13))에 의해 역전사시켰다. 태반 cDNA

샘플에 대해, 웰 당 대략 1개의 주형 분자에서 후속 디지털 PCR 증폭이 수행될 수 있도록 희석 최적을 수행하였다.

[0239] 12.765 디지털 어레이 (플루이다임)에 의해 바이오마크 시스템™ (플루이다임) 상에서 디지털 PCR을 수행하였다. 각각의 디지털 어레이는 12개의 샘플-검정 혼합물을 수용하기 위한 12개의 패널로 이루어져 있다. 각각의 패널을 7 nL 반응물/웰을 수행하기 위한 765개 웰로 추가 분할하였다. PLAC4 유전자 상의 rs8130833 SNP 영역을 겨냥한 프라이머 (5'-TTTGTATTGCAACACCATTTGG-3' (서열 14)) 및 앞서 기재된 유전자-특이적 역전사 프라이머에 의해 증폭시켰다. rs8130833 SNP의 2개의 대립유전자를 각각 표적으로 하는 2개의 대립유전자-특이적 헤르만 프로브를 설계하였다. 그들의 서열은 G 및 A 대립유전자에 대하여 각각 5'-(FAM)TCGTCGTCTAACTTG(MGBNFQ)-3' (서열 15) 및 5'-(VIC)ATTCGTCATCTAACTTG(MGBNFQ)-3' (서열 16)였다. 반응 부피 10 μL로 2X 헤르만 유니버셜 PCR 마스타 믹스를 사용하여 1개의 어레이 패널에 대한 반응을 설정하였다. 각각의 반응은 1X 헤르만 유니버셜 PCR 마스타 믹스, 각각의 프라이머 572 nM, 대립유전자-G-특이적 프로브 53.5 nM, 대립유전자-A-특이적 프로브 178.5 nM 및 cDNA 샘플 3.5 μL를 함유하였다. 각각의 태반 cDNA 샘플에 대해서는 1개의 반응 패널이 사용된 반면, 각각의 모체 혈장 샘플에 대해서는 12개의 패널이 사용되었다. 샘플-검정 혼합물을 나노플렉스(NanoFlex)™ IFC 제어기 (플루이다임)에 의해 디지털 어레이에 로딩하였다. 반응은 바이오마크™ 시스템에서 수행하였다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 40주기를 수행하였다.

[0240] 1개의 정배수체 및 2개의 T21 이형접합체 태반으로부터의 태반 RNA 샘플을 765-웰 반응 패널에서 분석하였다. 각각의 샘플에 대해, A 또는 G 대립유전자 (둘 다는 아님)에 대해 양성인 것을 포함하는 정보제공 웰의 수를 계수하였다. 모든 정보제공 웰 중 과다표출된 대립유전자의 비율 (P_i)을 결정하였다. 이를 실행의 웰 당 정확한 평균 참조 주형 농도 (m_i)에 적당한 SPRT 우선을 적용하여, 실험적으로 얻은 P_i이 정배수체를 지시하는지 T21 샘플을 지시하는지를 결정하였다. 도 15a에 나타낸 바와 같이, 상기 접근법을 사용하여 모든 RNA 샘플이 정확하게 분류되었다.

[0241] 정배수체 태아를 가진 4명의 여성 및 삼염색체 21 태아를 가진 1명의 여성으로부터 채취한 혈장 RNA 샘플을 추가로 시험하였다. 각각의 샘플을 12개의 765-웰 반응 패널, 즉 혈장 RNA 샘플 당 9180회의 반응으로 분석하였다. 도 15b는 상기 혈장 RNA 샘플의 12개 패널 각각에 대한 정보제공 웰의 수를 나타낸다. 표에 나타낸 바와 같이, 혈장 샘플 중 주형 농도는 임의의 1개의 반응 패널 중의 정보제공 웰의 수로는 SPRT 분류에 충분하지 않을 정도로 묽었다. 상기 샘플을 정배수체 샘플로 분류하기 전에, 2개의 반응 패널로부터의 정보제공 웰을 합해야 했다 (도 15c). 도 15c는 2 내지 12개의 패널로부터 통합된 데이타를 사용하여 모든 혈장 사례가 정확하게 분류될 수 있었음을 나타낸다.

[0242] 디지털 PCR을 수행하기 위한 메뉴얼 방법과 비교해서, 상기 미세유체-기반 방법은 훨씬 더 빠르고 덜 노동 집약적이었다. 전체 과정이 2시간 반 내에 완료될 수 있었다.

[0244] **B. 삼염색체 18의 출생 전 검출을 위한 디지털 RNA-SNP 분석**

[0245] 본 실시예에서는, 삼염색체 18 태아에 대한 다형성 대립유전자의 비율 불균형을 검출하기 위해, 염색체 18 상의 태반-발현 전사체인 세르핀 펩티다제 억제제 클레이드 B (오브알부민) 멤버 2 (SERPINB2) mRNA에 대한 디지털 PCR-기반 대립유전자 식별 검정을 사용하였다. QIAamp DNA 미니 키트 (독일 힐덴 소재의 키아겐(Qiagen)) 및 TRIzol 시약 (미국 캘리포니아주 칼즈배드 소재의 인비트로젠)을 제조업자의 프로토콜에 기재된 바와 같이 각각 사용하여, 태반 조직 샘플로부터의 DNA 및 RNA 추출을 수행하였다. 추출한 태반 RNA 샘플을 DNase I (인비트로젠) 처리하여 오염된 게놈 DNA를 제거하였다. 앞서 기재된 바와 같은 맥스어레이 컴팩트(MassARRAY Compact) (미국 샌디에고 소재의 시퀘놈(Sequenom))를 사용하여 호모지니스 맥스익스텐드(Homogeneous MassEXTEND) (hME) 검정에 의해 태반 조직 DNA 샘플에서 SERPINB2 유전자 상 rs6088 SNP의 유전자형 결정을 수행하였다.

[0246] 써모스크립트 역전사효소 (인비트로젠)를 사용하여 유전자-특이적 프라이머 5'-CGCAGACTTCTCACCAAACA-3' (서열 17) (미국 아이오와주 코럴빌 소재의 인테그레이티드 DNA 테크놀러지스)에 의해 태반 조직 RNA 샘플 상에서 SERPINB2 전사체에 대한 역전사를 수행하였다. 모든 cDNA 샘플을, 반응 웰 당 1개의 주형 분자의 평균 농도에서 후속 디지털 PCR 증폭이 수행될 수 있도록 하는 농도로 희석하였다. 헤르만 유니버셜 PCR 마스타 믹스 (미국 캘리포니아주 포스터 시티 소재의 어플라이드 바이오시스템즈) 및 바이오마크™ PCR 시약 (미국 랜프란시스코 소재의 플루이다임)을 사용하여 디지탈 PCR을

설정하였다. 전방향 프라이머 5'- CTCAGCTCTGCAATCAATGC-3' (서열 18) (인체크레이티드 DNA 테크놀러지스) 및 역방향 프라이머 (역전사에 사용된 유전자-특이적 프라이머와 동일한 것)는 600 nM의 농도로 사용하였다. SERPINB2 서열 상 rs6093 SNP의 A 또는 G 대립유전자를 표적으로 하는 2개의 태그만 프로브는 5'-(FAM)CCACAGGGAATTATTT (MGBNFQ)-3' (서열 19) 및 5'-(VIC)CCACAGGGATTATTT(MGBNFQ)-3' (서열 20) (어플라이드 바이오시스템즈)였다. FAM은 6-카르복시플루오레세인이고, MGBNFQ는 마이너 그루브 결합 비-형광 켄처이며, 각각 200 nM 및 500 nM의 농도로 사용하였다. 각각의 샘플-시약 혼합물을 나노플렉스™ IFC 제어기 (플루이다임)를 사용하여 바이오마크™ 12.765 디지탈 어레이 상의 765개 반응 웰에 분할하였다. 분할 후, 어레이를 열 종목 및 형광 검출을 위해 바이오마크™ 실시간 PCR 시스템 (플루이다임)에 넣었다. 반응을 50℃에서 2분 동안 개시하고, 95℃에서 5분 동안 지속한 후, 95℃에서 15초 및 59℃에서 1분의 40주기를 수행하였다. 증폭 후, 정보제공 웰 (A 또는 G 대립유전자 단독에 대해 양성인 것)의 수 및 대립유전자 둘 다에 대해 양성인 웰의 수를 계수하여 순차 확률비 검증 (SPRT) 분석에 제공하였다.

[0247] 이형접합 정배수체 태아에 대해, A 및 G 대립유전자는 태아 게놈에서 동등하게 표출되어야 하는 반면 (1:1), 삼염색체 18에 대해서는, 한 대립유전자의 추가 카피가 존재함으로써 태아 게놈 내에 2:1의 비율이 주어질 것이다. 상이한 샘플의 해석에 대한 일련의 SPRT 곡선을 생성시켰다. 이들 곡선은 분류에 필요한 정보제공 웰의 주어진 총 수 (x축)에 대한, 과다표출된 대립유전자에 대해 양성인 정보제공 웰의 예상 비율 P_1 (y축)을 나타낸다. 각각의 샘플에 대해, 실험적으로 유도된 P_1을 예상 P_1 값과 비교하였다. 상부 곡선 위쪽의 샘플은 삼염색체 18로 분류되고, 하부 곡선 아래쪽의 샘플은 정배수체로 분류되었다. 두 곡선 사이의 공간은 분류불가능한 영역이다.

[0248] 태아 삼염색체 18의 검출을 위한 디지탈 RNA-SNP 분석의 실행가능성을 SERPINB2 유전자 상 rs6098 SNP를 사용하여 증명하였다. 먼저 이형접합체 사례를 확인하기 위해서, 질량 분광법에 의해 정배수체 및 삼염색체 18 태아를 갖는 대상체로부터의 태반 조직 DNA 샘플의 유전자형을 결정하였다. 9개의 정배수체 및 3개의 삼염색체 18 이형접합체 태반을 확인하여 디지탈 RNA-SNP 분석에 제공하였다. 각각의 샘플에 대해 P_1 및 m_c을 계산하고, 상기 m_c값에 대한 SPRT 곡선을 질환 분류를 위해 사용하였다. 도 16a에 나타낸 바와 같이, 모든 샘플이 정확하게 분류되었다. 삼염색체 18 태반에 대한 P_1값은 분류불가능한 영역의 위쪽인 반면, 정배수체 태반에 대한 P_1값은 상기 영역의 아래쪽에 속하였다.

[0249] m_c = 0.1, 0.2 및 0.3을 기준으로 한 SPRT 곡선을 갖는 샘플을 도 16b에 나타내었다. 이들 데이타는 디지탈 RNA-SNP 방법이 삼염색체 18 임신에 대한 매우 유용한 진단 도구임을 시사한다. 2개의 곡선은 분류불가능한 영역에 대한 경계선을 나타낸다. 상부 곡선 위쪽의 데이타 지점을 갖는 샘플은 이수체로 분류되고, 하부 곡선 아래쪽의 데이타 지점을 갖는 샘플은 정배수체로 분류되었다.

[0251] C. 디지탈 RCD 분석

[0252] 본 실시예는 미세유체-기반 디지탈 PCR을 사용한 디지탈 RCD 분석의 성능을 증명한다. 플루이다임 바이오마크™ 시스템을 사용하는 이 접근법의 변형이 그에 제한되지 않는 예로서 본원에 예시되었다. 상기 시스템은 실행 1회당 9000회 초과의 디지탈 PCR을 수행할 수 있다.

[0253] 정배수체 또는 삼염색체 21 (T21) 태아를 가진 임신 여성으로부터 태반 조직, 모체 혈액 세포 및 양수 샘플을 얻었다. 10개의 T21 및 10개의 정배수체 사례로부터의 태반 DNA를 각각 동량의 정배수체 모체 혈액 세포 DNA와 혼합함으로써 20개의 50% DNA 혼합물을 제조하였다. 혼합물 샘플 중 정확한 태아 비율을 확실히 하기 위해, 추출한 DNA를 먼저 260 nm에서 광학 밀도 (OD) 측정에 의해 정량화하였다. 그 다음, 12.765 디지탈 어레이 (플루이다임)를 사용하여 바이오마크™ 시스템 (플루이다임)에 의해 이들을 디지탈 정량화하였다. 샘플의 정량화 결정은 참조 염색체에 대한 프로브를 사용한 것만 제외하고는 이하에 기재된 것과 동일하였다.

[0254] 50% DNA 혼합물 및 양수 샘플 내 염색체 양을, 염색체 1 상에 위치한 유전자좌에 대한 비-다형성 염색체 21 유전자좌의 디지탈 PCR 분석에 의해 결정하였다. 먼저 염색체 21 및 1 상의 1쌍의 파라로그 유전자좌의 101-bp 앰플리콘을 전방향 프라이머 5'-GTTGTTCTGCAAAAAACCTTCGA-3' (서열 21) 및 역방향 프라이머 5'-CTTGGCCAGAAATACTTCATTACCATAT-3' (서열 22)에 의해 동시 증폭시켰다. 염색체 21 및 1 파라로그 사이를 구별하도록 2개의 염색체-특이적 태그만 프로브를 설계하였는데, 그들의 서열은 각각

5'-(FAM)TACCTCCATAATGAGTAAA(MGBNFQ)-3' (서열 23) 및 5'-(VIC)CGTACCTCTGTAATGTGTAA(MGBNFQ)-3' (서열 24)였다. 본원에서 파라로그 유전자좌는 단지 예로서 사용되었다. 용어하면, 비-파라로그 유전자좌도 이러한 분석에 대해 사용될 수 있다.

[0255] 상염색체 18 (T18)을 검출하기 위한 디지탈 RCD 접근법의 용도를 증명하기 위해, 염색체 21 및 18 상의 파라로그 서열을 표적으로 하는 또다른 검정을 설계하였다. 먼저 염색체 21 및 18 상의 파라로그 유전자좌의 128-bp 앰플리콘을 전방향 프라이머 5'-GTACAGAAACCACAAACTGATCGC-3' (서열 25) 및 역방향 프라이머 5'-GTCCAGGCTGTGGGCCT-3' (서열 26)에 의해 동시 증폭시켰다. 염색체 21 및 18 파라로그 사이를 구별하도록 2개의 염색체-특이적 테크만 프로브를 설계하였는데, 그들의 서열은 각각 5'-(FAM)AAGAGGCGAGGCAA(MGBNFQ)-3' (서열 27) 및 5'-(VIC)AAGAGGACAGGCAAC(MGBNFQ)-3' (서열 28)였다. 본원에서 파라로그 유전자좌는 단지 예로서 사용되었다. 용어하면, 비-파라로그 유전자좌도 이러한 분석에 대해 사용될 수 있다.

[0256] 모든 실험은 12.765 디지탈 어레이 (폴루이다임)를 사용하여 바이오마크™ 시스템 (폴루이다임) 상에서 수행하였다. 반응 부피 10 μL로 2X 테크만 유니버셜 PCR 마스타 믹스 (어플라이드 바이오시스템즈)를 사용하여 1개의 웰에 대한 반응을 설정하였다. 각각의 반응은 1X 테크만 유니버셜 PCR 마스타 믹스, 각각의 프라이머 900 nM, 각각의 프로브 125 nM 및 50% 태반/모체 혈액 세포 DNA 샘플 3.5 μL를 함유하였다. 샘플/검정 혼합물을 나노플렉스™ IFC 제어기 (폴루이다임)에 의해 디지탈 어레이에 로딩하였다. 검출을 위해 바이오마크™ 시스템 상에서 반응을 수행하였다. 반응을 50℃에서 2분 동안 개시하고, 이어서 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 40주기를 수행하였다.

[0257] 정배수체 및 T21 50% 태반/모체 혈액 세포 DNA 샘플을 디지탈 어레이 상에서 chr21/chr1 검정에 의해 분석하였다. 각각의 샘플에 대해, chr21 또는 chr1 마커 단독 (둘 다는 아님)에 대해 양성인 웰을 포함하는 정보제공 웰의 수를 계수하였다. 모든 정보제공 웰 중 파다표출된 마커의 비율 (P_r)을 결정하였다. 디지탈 PCR 패널 중 어느 하나에 대한 웰 당 경험적 평균 참조 주형 농도 (m_r)에 적당한 SPRT 곡선을 적용하여, 실험적으로 얻은 P_r이 정배수체 샘플을 지시하는지 T21 샘플을 지시하는지를 결정하였다. 결정이 이루어질 수 있을 때까지, 분류되지 않은 채로 남아있는 샘플에 대한 여분의 패널로부터 데이타를 종합하였다. 도 17에 나타낸 바와 같이, 모든 50% 태반/모체 혈액 세포 DNA 샘플이 상기 접근법을 사용하여 필요한 1 내지 4개 패널 범위의 데이타에 의해 정확하게 분류되었다. 도 18에 나타낸 바와 같이, 정확한 분류를 위한 결정 경계선을 나타내기 위해 SPRT 곡선을 또한 플롯팅하였다.

[0258] 정배수체 태아를 가진 23명의 여성 및 T21 태아를 가진 6명의 여성으로부터 얻은 양수 샘플에 대한 RCD 분석을 추가로 적용하였다. 각각의 샘플을 단일 765-웰 반응 패널에서 chr21/chr1 검정에 의해 분석하였다. 도 19는 SPRT 분류 개요를 나타낸다. 도 19에 나타낸 바와 같이, 모든 29개 샘플이 정확하게 분류되었다. 따라서, 디지탈 RCD 방법은 출생 전 진단에 사용되는 다양한 종류의 샘플 유형, 예컨대 양수 및 융모막 융모 생검에서 초위성체 (문헌 [Levett LJ, et al. A large-scale evaluation of amnio-PCR for the rapid prenatal diagnosis of fetal trisomy. Ultrasound Obstet Gynecol 2001; 17: 115-8]) 또는 단일 뉴클레오티드 다형성 (SNP) (문헌 [Tsui NB, et al. Detection of trisomy 21 by quantitative mass spectrometric analysis of single-nucleotide polymorphisms. Clin Chem 2005; 51: 2358-62]) 마커 또는 실시간 비-디지탈 PCR (문헌 [Zimmermann B, et al. Novel real-time quantitative PCR test for trisomy 21. Clin Chem 2002; 48: 362-3])을 사용하는 상염색체의 검출에 대한 대안적인 접근법이다.

[0259] T18 사례를 검출하기 위한 시도로, 3개의 정배수체 및 5개의 T18 태반 DNA 샘플 상에 chr21/chr18 검정을 적용하였다. 모든 정보제공 웰 중 파다표출된 마커의 비율 (P_r)을 계산하였다. 이들 모두는 1개의 T18 사례가 정배수체로 잘못 분류된 것을 제외하고는 정확하게 분류되었다. 결과를 도 20에 요약하였다.

[0261] V. 혈장 분획물 샘플에 대한 멀티플렉스 디지탈 RCD 검정의 이용

[0262] 모체 혈장의 단위 부피 당 존재하는 혈장 DNA 분자의 수는 제한적이다 (문헌 [Lo YMD, et al. 1998 Am J Hum Genet 62, 768-7758]). 예를 들어, 임신 초기에 상염색체 유전자좌, β-글로빈 유전자의 중앙 모체 혈장 농도는 태아 및 모계 모두에 기인하여 286개 카피/mL인 것으로 밝혀진 바 있다 (문헌 [Lo YMD, et al. 1998 Am J

Hum Genet 62, 763-7758]). 7,680개의 분자를 포획하기 위해서, 대략 8 mL의 모체 혈장으로부터 추출된 DNA가 필요할 것이다. 약 15 mL의 모체 혈액으로부터 수득가능한 상기 부피의 혈장은 통상적인 실행의 한계이다. 그러나, 본 발명자들은 chr21 및 참조 염색체 표적의 다중 세트를 디지탈 RCD 분석을 위해 합할 수 있는 것으로 생각한다. 5쌍의 chr21 및 참조 염색체 표적에 대해, 단지 1.6 mL의 모체 혈장이 분석에 필요한 주형 분자의 수를 제공하기 위해 필요할 것이다. 멀티플렉스 단일 분자 PCR을 수행할 수 있었다. 이러한 멀티플렉스 단일 분자 분석의 견고성은 단일 분자 일배체형화에 대해 이미 입증된 바 있다 (문헌 [Ding, C. and Cantor, CR. 2003 Proc Natl Acad Sci U S A 100, 7449-7453]).

[0263] 임베에서, 태반 조직 및 모체 혈액 세포 샘플을 정배수체 또는 삼염색체 21 (T21) 태아를 가진 임신 여성으로부터 수득하였다. 5개의 정배수체 및 5개의 T21 태반 DNA 샘플을 각각 동일한 비율의 모체 혈액 세포 DNA와 혼합하여 50% 태아 DNA를 갖는 혈장 샘플을 모방한 10개의 DNA 혼합물을 생성하였다. 혼합을 샘플 중 정확한 태아 비율을 확실하게 하기 위해, 추출된 DNA를 먼저 260 nm에서의 광학 밀도 (OD) 측정으로 정량화하였다. 이어서, 이를 384-웰 포맷에서 실시간 PCR에 의해 디지탈 방식으로 정량화하였다. 샘플을 정량화하기 위한 검정은 상기 디지탈 RCD 분석의 예에 기재된 바와 동일하였다.

[0264] 50% 믹스 내 염색체 양을 염색체 1 상에 위치한 유전자좌에 대한 비-다형성 염색체 21 유전자좌의 디지탈 PCR 분석으로 결정하였다. 상기 방법은 디지탈 상대적 염색체 양 (RCD) 분석으로 칭해진다. 염색체 21 및 1 상의 한 쌍의 파라로그 유전자좌의 121-bp 앰플리콘 (각각의 프라이머 상에 10-mer 포함)을 전방향 프라이머 5'-ACGTTGGATGGTTGTTCTGCAAAAACCTTCGA-3' (서열 29) 및 역방향 프라이머 5'-ACGTTGGATGCTTGGCCAGAAATACTTCATTACCATAT-3' (서열 30)를 사용하여 동시 증폭시켰다. 염색체 21 및 염색체 1 사이의 염기 차이를 표적으로 한 연장 프라이머를 고안하였고, 그의 서열은 5'-CTCATCCTCACTTCGTACCTC-3' (서열 31)이다.

[0265] T21 사례를 검출하는 멀티플렉싱 디지탈 PCR 검정의 유용성을 입증하기 위해서, 염색체 21 및 18 상의 파라로그 서열을 표적으로 한 또다른 디지탈 RCD 검정을 고안하였다. 염색체 21 및 18 상의 파라로그 유전자좌의 148-bp 앰플리콘 (각각의 프라이머 상에 10-mer 포함)을 전방향 프라이머 5'- ACGTTGGATGGTACAGAACCACAAACTGATCGG-3' (서열 32) 및 역방향 프라이머 5'-ACGTTGGATGCTCC AGGCTGTGGGCCT-3' (서열 33)를 사용하여 동시 증폭시켰다. 염색체 21 및 염색체 18 사이의 염기 차이를 표적으로 한 연장 프라이머를 고안하였고, 그의 서열은 5'-ACAAAGGGGGAAGAGG-3' (서열 34)이다.

[0266] 멀티플렉스 디지탈 RCD 분석은 프라이머 연장 프로토콜을 이용하여 수행되었다. PCR 반응은 겐앰프(GeneAmp) PCR 코어 시약 키트 (어플라이드 바이오시스템즈)를 사용하여 반응 부피 5 μL로 설정하였다. 각각의 반응물은 1X 완충액 II, 2 mM $MgCl_2$, 200 μM dNTP 믹스, 0.2 U 앰플리태크 골드(AmpliTaq Gold), 각각 300 nM의 4종의 프라이머 및 50% DNA 믹스를 함유하였다. 상기 검정/샘플 혼합물을 384-웰 PCR 플레이트에 분배하고, 50℃에서 2분 동안 반응을 개시한 후, 95℃에서 10분 동안 진행하고, 95℃에서 15초 및 57℃에서 1분의 40회 주기를 수행하였다.

[0267] PCR 생성물을 쉬림프 알칼린 포스파타제 (SAP) 처리하여 혼입되지 않은 dNTP를 제거하였다. 상기 혼합물을 37℃에서 40분 동안, 이어서 85℃에서 5분 동안 인큐베이션하였다. 이어서, 프라이머 연장 반응을 수행하였다. 간단히 말해서, 연장 칵테일 중 chr21/chr1 검정으로부터의 771 nM의 연장 프라이머, chr21/chr18 검정으로부터의 1.54 μM의 연장 프라이머, 0.67 U 써모시퀘나제(Thermosequenase) (시퀘놈), 및 각각 64 μM의 ddCTP, ddGTP, dATP 및 dTTP를 SAP-처리된 PCR 생성물에 첨가하였다. 반응 조건은 94℃에서 2분, 이어서 94℃에서 5초, 50℃에서 5초 및 72℃에서 5초의 80회 주기였다. 16 μL의 물 및 3 mg의 물린 레진(Clean Resin) (시퀘놈)을 최종 처리(clean up)를 위해 연장 생성물에 첨가하였다. 상기 혼합물을 회전기에서 20 내지 30분 동안 혼합한 후, 361 g에서 5분 동안 원심 분리단계를 수행하였다. 15 내지 25 nL의 최종 생성물을 매스어레이 나노 분산기 S (시퀘놈)로 스펙트로칩(SpectroCHIP) (시퀘놈) 상에 분배하였다. 스펙트로칩으로부터의 데이타 획득을 매스어레이 분석기 컴팩트 질량 분광기 (시퀘놈)에서 수행하였다. 분석을 위해 질량 데이타를 매스어레이 타이퍼(MassARRAY Typer) (시퀘놈) 소프트웨어로 임포트하였다.

[0268] 5개의 정배수체 및 5개의 T21 50% 태반/모체 DNA 샘플을 이중 RCD 검정으로 분석하였다. 각 샘플에 대해, 오직 chr21 또는 chr1 또는 chr18 마커에만 양성인 웰을 포함하여 각각의 검정으로부터 정보제공 웰의 수를 계수하였다. 전체 정보제공 웰 중 chr21 마커의 비율 (P_i)을 각각의 RCD 검정에 대해 개별적으로 계산하였다. 이어서, 순차 확률비 검증 (SPRT)을 적용하여 P_i이 정배수체 샘플을 나타내는 지 또는 T21 샘플을 나타내는 지 결

정하였다. 이렇게 하여, 각각의 플레이트를 2회 제수함으로써 필요한 웰의 수를 감소시켰다.

[0269] chr21/chr1 결정을 통상 먼저 적용하였다. 샘플이 이분류 상태로 있는 경우, chr21/chr18 결정으로부터의 값을 추가 계산에 포함시킬 것이다. 여분의 플레이트를 결정이 이루어질 수 있을 때까지 이분류 상태로 있는 샘플에 사용하였다. 도 21에 나타낸 바와 같이, 모든 정배수체 50% 믹스 샘플을 단일 384-웰 플레이트를 사용하여 정확히 분류하였다. 여러 T21 사례는 정확한 분류를 위해 2개 이상의 플레이트를 필요로 하였다. 하나의 결정만을 이용한 경우, 분류 달성시 필요한 정보제공 웰의 수에 이르기 위해서 더 많은 플레이트가 필요할 것이다. 예를 들어, RCD 결정 중 어느 하나를 단독 이용한 경우에는 사례 KO230에 대한 데이타가 분류될 수 없었다. 그러나, 2개의 결정으로부터의 데이타를 함께 경우에는 정확한 분류가 이루어졌다. 이중 RCD 결정을 이용하지 않은 경우, 추가의 분석 플레이트가 필요할 것이다. 본 발명자들은 보다 높은 수준의 멀티플렉싱 결정으로 웰 수의 추가 감소를 예상할 것이다.

[0270] 또다른 예에서, 본 발명자들은 염색체 21, 및 다른 비-염색체-21 상염색체 상에 위치한 그의 상응하는 파라로그 파트너에 대한 4종의 상이한 샘플라존을 표적으로 하는 4-플렉스 결정을 개발하였다. 이러한 4-플렉스 결정은 정배수체 및 삼염색체 21 임신 여성으로부터의 샘플의 디지털 RCD 분석, 이어서 SMRT 분류에 사용되었다. 태반 샘플로부터의 DNA 추출은 리아젠프 조직 키트 (독일 힐덴 소재의 퀴아겐)를 사용하여 수행되었다.

[0271] 상기 연구에 사용된 모든 태반 및 모체 연막 DNA 샘플을 먼저 나노드롭 분광광도계 (미국 델라웨어주 윌밍턴 소재의 나노드롭 테크놀로지)로 정량화하였다. DNA 농도를 6.6 pg/세포의 전환을 이용하여 게놈 당량 (GE)/μL로 전환하였다. 웰 당 대략 1개의 주형에 해당하는 DNA의 양을 DNA 샘플의 계열 희석으로 결정하였다. 이러한 조건하에, 본 발명자들은 대략 37%의 웰이 음성 증폭을 나타낼 것으로 예상할 것이다. 멀티플렉스 디지탈 RCD 분석에서, 4 세트의 파라로그 서열 표적을 선택하였다: 염색체 21 및 1 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGTTGATGAAGTCTCATCTCTACTTCG-3' (서열 35) 및 역방향 프라이머 5'-ACGTTGGATGCAATAAGCTTGGCCAGAAATACT-3' (서열 36)을 사용하여 동시 증폭시켜 81 bp의 앰플리콘을 생성하였다. 염색체 21 및 7 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGGAATTTAAGCTAAATCAGCCTGAACTG-3' (서열 37) 및 역방향 프라이머 5'-ACGTTGGATGGTTTCTCATGTTCATCGTAGGCTTAT-3' (서열 38)를 사용하여 동시 증폭시켜 82 bp의 앰플리콘을 생성하였다. 염색체 21 및 2 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGTCAGGCAGGGTTCTATGCAG-3' (서열 39) 및 역방향 프라이머 5'-ACGTTGGATGAGCGGCTTCCTGGCTCTTG-3' (서열 40)를 사용하여 동시 증폭시켜 101 bp의 앰플리콘을 생성하였다. 염색체 21 및 6 상의 파라로그 유전자좌를 전방향 프라이머 5'-ACGTTGGATGGCTCGTCTCAGGCTCGTAGTT-3' (서열 41) 및 역방향 프라이머 5'-ACGTTGGATGTTTCTTCGAGCCCTTCTTGG-3' (서열 42)를 사용하여 동시 증폭시켜 102 bp의 앰플리콘을 생성하였다. 각각의 반응물은 10X 완충액 1l (어플라이드 바이오시스템즈), $MgCl_2$ 및 각각 100 nM의 프라이머를 함유하였다.

총 반응 부피는 5 μL/웰이었다. 95℃에서 5분 동안 반응을 개시한 후, 25℃에서 30초, 62℃에서 20초 및 72℃에서 30초의 45회 주기를 수행하고, 72℃에서 7분 동안 최종 연장을 수행하였다. 모든 통상적인 PCR 증폭은 진앰프 PCR 시스템 9700 (어플라이드 바이오시스템즈) 상에서 수행되었다. 혼입되지 않은 뉴클레오티드를 쉬림프 알칼린 포스파타제 (SAP) 처리로 비활성화시켰다. 각각의 반응물은 10X SAP 완충액 (시퀘놈) 및 SAP 효소 (시퀘놈)을 함유하였다. 2 μL의 SAP 믹스를 각각의 PCR에 첨가하였다. SAP 반응물을 37℃에서 40분 동안 및 85℃에서 5분 동안 인큐베이션하였다. SAP 처리 후, iPLEX 골드(Gold) 키트 (시퀘놈)를 사용하여 PCR 생성물 상에서 프라이머 연장 반응을 수행하였다. 염색체 21 및 1 상의 파라로그 유전자좌에 대한 파라로그 서열 비스해저 (PSM)는 연장 프라이머 5'-GTCTCATCTCTACTTCGTACCTC-3' (서열 43)에 의해 검사되었다. 염색체 21 및 7 상의 파라로그 유전자좌에 대한 PSM은 연장 프라이머 5'-TTTTACGCTGTCCCCATTT-3' (서열 44)에 의해 검사되었다. 염색체 21 및 2 상의 파라로그 유전자좌에 대한 PSM은 연장 프라이머 5'-GGTCTATGCACGAGCCGAC-3' (서열 45)에 의해 검사되었다. 염색체 21 및 6 상의 파라로그 유전자좌에 대한 PSM은 연장 프라이머 5'-TGGGCGCGGGAGCGGACTTCGCTGG-3' (서열 46)에 의해 검사되었다. 각각의 반응물은 10X iPLEX 완충액 (시퀘놈), iPLEX 종결 믹스 (시퀘놈), iPLEX 효소 (시퀘놈), 및 1.63 μM로 사용된 염색체 21 및 6에 대한 PSM용 연장 프라이머를 제외하고는 각각 343 nM의 연장 프라이머를 함유하였다. 2 μL의 iPLEX 믹스를 5 μL의 PCR 생성물에 첨가하였다. 200-쇼트-사이클 프로그램에 따라 iPLEX 반응을 순환시켰다. 간단히 말해서, 샘플을 먼저 94℃에서 35초 동안 변성시킨 후, 52℃에서 5초 동안 어닐링하고, 80℃에서 5초 동안 연장하였다. 어닐링 및 연장 주기를 총 5회의 주기 동안 4회 이상 반복하고, 이어서 94℃, 5초의 변성 단계로 역 순환시킨 후, 다시 5회 주기의 어닐링 및 연장 루프를 수행하였다. 1회의 변성 단계와 함께 5회의 어닐링 및 연장 주기를 30회 반복하여 총 40회에 이르게 하였다. 72℃에서 3분 동안 최종 연장을 수행하였다. iPLEX 반응 생성물을 16 μL의 물로 희석하고, 각각의 PCR에 대해 6 mg의 수지로 탈염시켰다. 384-웰 플레이트를 1600 g에서 3분 동안 원심분리한

후, 스펙트로칩 (시퀘놈) 및 에트릭스-보조 레이저 탈착/이온화 비행시간형 (MALDI-TOF) 질량 분광법 MS 분석기 (시퀘놈) 상에 분배하였다.

[0272] 염색체 21에 대해서만 또는 4개의 검정 각각에 대한 참조 염색체에 대해서만 양성인 웰의 수를 독립적으로 기록하였다. 각각의 검정에 있어서, 염색체 21 및 참조 염색체에 대한 푸아송 교정된 분자 수를 계산하였다. 염색체 21에 대한 푸아송 교정된 분자 수의 합 뿐만 아니라 분석 4개의 검정으로부터의 푸아송 교정된 참조 염색체 수의 합을 계산하였고, 이는 4-플렉스 검정에 대한 정보제공 계수로 간주되었다. P_i 값은 4-플렉스 검정에 대한 염색체 21 계수를 4-플렉스 검정에 대한 염색체 21 및 참조 염색체 계수의 합으로 나눈 값이었다. 실험적으로 유도된 P_i 값을 SPRT 분석에 적용하였다. SPRT에 의해 질환 분류가 가능해질 때까지 1개 이상의 384-웰 플레이트를 분석할 것이다. 총 2개의 50% 정배수체 대한 계놈 DNA/50% 모체 인막 DNA 믹스 및 2개의 50% 삼염색체 21 대한 계놈 DNA/50% 모체 인막 DNA 믹스를 분석하였다.

[0273] 실험적으로 유도된 P_i 값을 귀무가설 또는 대립가설 시험시의 예상 P_i 값과 비교할 것이다. 다르게는, 주어진 수의 정보제공 계수에 대한 P_i이 질환 분류에 필요한 통계적 신뢰 수준에 도달하지 못한 경우, 귀무가설 또는 대립가설은 모두 수용될 수 없을 것이다. 이러한 경우는 보다 많은 데이타가 이용가능해질 때까지 분류불가능인 것으로 간주되었다.

[0274] 각 샘플의 결과 및 SPRT 분류를 도 22a 및 22b에 표로 작성하였다. 2개의 정배수체 샘플은 SPRT 분류에 도달할 수 있기 전에 2개 및 5개의 384-웰 멀티플렉스 디지탈 RCD 분석을 필요로 하였다. 4-플렉스 검정의 개별 구성으로부터의 데이타는 어떤 것도 SPRT에 의한 질환 분류가 가능하지 않았다. 삼염색체 21 샘플은 둘 다 단지 하나의 384-웰 멀티플렉스 디지탈 RCD 분석으로 각각 정확히 분류되었다. 유사하게, 4-플렉스 검정의 개별 구성으로부터의 데이타는 어떤 것도 SPRT에 의한 질환 분류가 가능하지 않았다. 그러나, 4-플렉스 검정으로부터의 혼합 계수는 정확한 SPRT 분류를 가능하게 하였다. 이들 데이타는, 멀티플렉스 디지탈 RCD의 이용에 의해 정보제공 계수의 유효 수가 단일-플렉스 디지탈 RCD 검정의 이용과 비교하여 주어진 수의 수행된 디지탈 PCR 분석에 대해 상당히 증가하였음을 예증하였다.

[0276] VI. 디지탈 후성 상대적 염색체 양의 사용

[0277] 여기서, 본 발명자들은, 염색체 이수성에 관여하는 염색체 (즉, 삼염색체 21에서 염색체 21) 및 참조 염색체 상의 태아-특이적 DNA 메틸화 패턴 또는 다른 후성 변화를 나타내는 후성 마커를 디지탈 PCR 분석에 적용하는, 소위 디지탈 후성 상대적 염색체 양 (디지탈 ERCD)이라 불리는 절근법을 약술한다. 정상 태아를 가진 임신 여성으로부터 추출된 혈장 DNA 내 참조 염색체 후성 마커에 대해 양성인 웰 수에 대한 염색체 21 후성 마커에 대해 양성인 웰 수의 비율은 참조 범위를 제공할 것이다. 상기 비율은 태아가 삼염색체 21을 갖는 경우 증가할 것으로 예상될 것이다. 하나 초과의 염색체 21 마커 및 하나 초과의 참조 염색체 마커를 본 분석에서 사용할 수 있다는 것이 당업자에게 명백하다.

[0278] 태아 (태반)-특이적 메틸화 패턴을 나타내는 염색체 21 상의 유전자의 일례는 홀로카르복실라제 합성효소 (Holocarboxylase synthetase, HLCS) 유전자이다. HLCS는 태반에서 과-메틸화되지만, 모체 혈액 세포에서는 저-메틸화되며, 이는 본원에 참조로 포함된 미국 특허 출원 제11/784480호에 포함된다. 태아 (태반)-특이적 메틸화 패턴을 나타내는 참조 염색체 상의 유전자의 일례는 염색체 3 [3p] 상의 RASSF1A 유전자이다. RASSF1A는 태반에서 과-메틸화되지만, 모체 혈액 세포에서는 저-메틸화되며, 본원에 참조로 포함된 미국 특허 출원 제11/784501호를 참조한다.

[0279] 모체 혈장을 이용한 태아 내 삼염색체 21의 디지탈 PCR 검출에 대한 저-메틸화된 HLCS 및 과-메틸화된 RASSF1A의 적용에서, 모체 말초 혈액을 먼저 수집한다. 이어서, 상기 혈액을 원심분리하고, 혈장을 수집한다. 이어서, 혈장으로부터의 DNA을 당업계에 널리 공지된 기술, 예컨대 퀴아엠프 블러드(Blood) 키트 (퀴아겐)을 사용하여 추출한다. 이어서, 혈장 DNA를 1종 이상의 메틸화-감수성 제한 효소, 예컨대 HpaII 및 BstUI를 사용하여 소화시킨다. 이들 메틸화-감수성 제한 효소는 태아의 과-메틸화된 서열을 무손상 상태로 두는 반면, 이들 유전자의 모체, 비-메틸화된 버전은 절단할 것이다. 이어서, 소화된 혈장 DNA 샘플을, 제한 효소 처리되었지만 무손상인 HLCS 또는 RASSF1A 서열의 분자가 반응 웰 당 평균 대략 0.2 내지 1개 집중될 정도로 희석한다. 2개의 실시간 PCR 시스템을 사용하여 희석된 DNA를 증폭시킬 것이며: 하나는 2개의 프라이머, 및 HLCS 유전자에 특이적인 1개의 태그된 프로브 (서열이 비-메틸화된 경우 제한 효소(들)에 의해 절단될 영역 포함)을 사용하고; RASSF1A 유전자에 관한 다른 하나는 유사하게 2개의 프라이머 및 1개의 태그된 프로브를 사용한다. 추가의 RASSF1A 프라이머/프로브 세트와 관련하여, 일례가 문헌 [Chan et al 2006, Clin Chem 52, 2211-2218]에 기재

되어 있다. HLCS 및 RASSF1A 표적에 관한 태크만 프로브는 상이한 형광 리포터, 예컨대 각각 FAM 및 VIC을 가질 것이다. 이어서, 384-웰 플레이트를 사용하여 디지탈 PCR 실험을 수행한다. HLCS에 대해서만 양성으로 스코어링된 웰의 수 및 RASSF1A에 대해서만 양성으로 스코어링된 웰의 수를 계수할 것이며, 이들 계수의 비율을 구할 것이다. HLCS:RASSF1A 비율은 정상 정배수체 태아를 가진 임신 여성과 비교하여 삼염색체 21 태아를 가진 임신 여성으로부터 얻은 모체 혈장에 대해 보다 높을 것으로 예상될 것이다. 차다표율의 정도는 디지탈 PCR 수행시 웰 당 평균 핵산 주형 농도에 따라 달라질 것이다.

[0280] 이들 결과를 스코어링 하는 다른 방법, 예를 들어 RASSF1A에 대해 동시에 양성인 것에 관계 없이 HLCS에 대해 양성인 웰의 수를 계수하고, 반대로 HLCS에 대해 동시에 양성인 것에 관계 없이 RASSF1A에 대해 양성인 웰의 수를 계수하는 방법이 가능할 것이다. 또한, 비율을 계산하는 대신 HLCS 및 RASSF1A 계수의 총 수 또는 이들의 차이를 이용하여 태아의 삼염색체 21 상태를 나타낼 수 있었다.

[0281] 플레이트에서 디지탈 PCR을 수행하는 것 외에도, 디지탈 PCR의 다양한 변형, 예를 들어 미세유체 칩, 나노리터 PCR 마이크로플레이트 시스템, 에멀젼 PCR, 롤로니 PCR, 롤링-써클 증폭법, 프라이머 연장법 및 질량 분광법 등을 이용할 수 있다는 것이 당업자에게 명백할 것이다. 디지탈 PCR의 이러한 변형은 단지 예로서 제시되며, 제한적이지 않다.

[0282] 실시간 PCR 외에도, 질량 분광법과 같은 방법을 이용하여 디지탈 PCR 결과를 스코어링할 수 있다는 것이 당업자에게 명백할 것이다.

[0283] HLCS 및 RASSF1A의 태아 및 모체 버전을 식별하기 위해 메틸화-감수성 제한 효소를 사용하는 것 외에도, 메틸화 상태를 확인하기 위한 다른 방법, 예를 들어 바이설파이트 변형법, 메틸화-특이적 PCR, 메틸화된 시토신에 대한 항체를 사용하는 면역침강법, 질량 분광법 등을 적용할 수도 있음이 당업자에게 명백할 것이다.

[0284] 또한, 본 특허 출원의 상기 예 및 기타 예에서 설명된 접근법을 모체 뇨, 양수, 자궁경관 세척액, 융모막 융모, 모체 타액 등을 비롯한 태아 DNA가 발견될 수 있는 기타 체액에서도 사용할 수 있음이 당업자에게 명백할 것이다.

[0286] VII. 에멀젼 PCR 및 기타 전략을 이용한 대규모 병렬 게놈 서열분석

[0287] 여기서, 본 발명자들은, 핵산 분자의 디지탈 판독이 모체 혈장 중 태아 염색체의 이수성, 예를 들어 삼염색체 21을 검출하기 위해 사용될 수 있기 때문에 핵산 분자의 디지탈 판독의 또다른 예를 기재할 것이다. 태아 염색체 이수성은 염색체 또는 염색체 영역의 비정상적 용량(들)으로부터 기인한다. 비침습성 산전법이 감수성 및 특이성이 높기 때문에 오진을 최소화하기에 바람직하다. 그러나, 태아 DNA는 낮은 절대 농도로 존재하고, 모체 혈장 및 혈청 중의 모든 DNA 서열에 대해 소량의 부분으로 나타난다. 따라서, 디지탈 PCR 샘플링 표적화에 특이적인 유전자좌의 갯수는 동일한 생물학적 표본 내에서 무한히 증가될 수 없다. 따라서, 많은 세트의 특이적 표적 유전자좌를 분석함으로써, 디지탈 PCR 샘플링 수행 횟수를 증가시키지 않으면서도 표본으로부터 얻을 수 있는 데이타의 양을 증가시킬 수 있다.

[0288] 따라서, 본 실시양태는 모체 혈장 핵산을 함유한 생물학적 샘플 중에 소량의 검단으로서 존재하는 제한된 양의 태아 핵산으로부터 추론될 수 있는 유전 정보의 양을 최대화함으로써 태아 염색체의 이수성을 비-침습적으로 검출한다. 한 측면에서, 얻어진 유전 정보의 양은 정확한 진단을 위해서는 충분하지만, 요구되는 생물학적 샘플의 투입량 및 비용을 포괄할 정도로 지나치게 과량인 것은 아니다.

[0289] 454 플랫폼 (로슈) (문헌 [Margulies, M. et al. 2005 Nature 437, 376-380]), 일루미나 게놈 분석기 (또는 솔렉사 플랫폼) 또는 솔리드 시스템 (어플라이드 바이오시스템즈) 또는 헬리코스 트루 싱글 몰레큘 DNA 서열분석 기술 (문헌 [Harris TD et al. 2008 Science, 320, 106-109]), 퍼시픽 바이오사이언시즈의 단일 분자, 실시간 (SMRT™) 기술, 및 나노기공 서열분석 (문헌 [Soni GV and Meller A. 2007 Clin Chem 53: 1996-2001]) 상에서 달성가능한 것과 같은 대규모 병렬 서열분석은 고도로 다중화된 표본으로부터 단리된 많은 핵산 분자를 병렬 방식으로 서열분석한다 (문헌 [Dear Brief Funct Genomic Proteomic 2003; 1: 397-416]). 이들 각각의 플랫폼은, 클로닝에 의해 확장된 또는 비-증폭된 핵산 집단의 단일 분자를 서열분석한다.

[0290] 대다수의 서열분석 판독이, 수십만 내지 수백만 또는 더 가능하게는 수백억 정도로, 각각의 진행에서 각각의 샘플로부터 생성되기 때문에, 얻어진 서열분석된 판독은 원래 표본 중 핵산 종 혼합물의 대표적인 프로파일을 형성한다. 예를 들어, 서열분석된 판독의 일배체형, 전사체 및 메틸화 프로파일은 원래 표본의 그것과 유사하다 (문헌 [Brenner et al Nat Biotech 2000; 18: 630-634]; [Taylor et al Cancer Res 2007; 67: 8511-8518]). 각각의 표본으로부터 서열을 대량 샘플링하기 때문에, 핵산 풀(pool)의 서열분석으로부터 몇배의

적용범위 또는 고 충복성으로 생성된 것과 같은 동일한 서열의 수는 또한 원래 샘플 중 특정 핵산 종의 수 또는 유전자의 수에 대한 좋은 정량적 표출이다.

[0291] 한 실시양태에서, 랜덤 서열분석은 임신 여성의 혈장 중에 존재하는 DNA 절편에 대해 수행되고, 분해 태아 또는 모계로부터 유래되는 게놈 서열을 얻는다. 랜덤 서열분석은 생물학적 샘플에 존재하는 핵산 분자의 랜덤 부분을 샘플링 (서열분석)하는 것을 포함한다. 서열분석이 랜덤하게 이루어지기 때문에, 핵산 분자 (및 따라서 게놈)의 상이한 하위세트 (분획)를 각각의 분석에서 서열분석할 수 있다. 본 실시양태는 이러한 하위세트가 샘플마다 및 심지어 동일한 샘플을 사용하여 이루어질 수 있는 분석마다 다양한 경우라 해도 작용할 것이다. 분획의 예는 게놈의 약 0.1%, 0.5% 또는 1%이다. 다른 실시양태에서, 분획은 이들 값 중 적어도 하나이다.

[0292] 이어서, 바이오인포메틱스 절차를 이용하여 이들 DNA 서열 각각을 인간 게놈에 위치시킬 수 있다. 상기 서열의 비율은 이들이 인간 게놈의 반복 영역 또는 개인간 변이, 예를 들어 카피수 변이(copy number variation)에 적용되는 영역에 존재하기 때문에 추후 분석으로부터 감산될 수 있다. 이에 따라, 본 실시양태 및 하나 이상의 다른 염색체의 양을 측정할 수 있다.

[0293] 한 실시양태에서, 염색체 이수성, 예를 들어 염색체 21 또는 염색체 18 또는 염색체 13에 잠재적으로 포함되는 염색체의 파라미터 (예를 들어, 분획 표출)는 예를 들어 바이오인포메틱스 절차의 결과로부터 계산될 수 있다. 분획 표출은 모든 서열 (예를 들어, 모든 염색체의 일부 측정치) 또는 특정 하위세트의 염색체 (예를 들어, 서열된 것과는 다른 염색체)의 양을 기준으로 얻어질 수 있다.

[0294] 이후, 상기 분획 표출은 정상 태아 (즉, 정배수체)를 포함하는 임신에서 확립된 참조 범위와 비교한다. 일부 변형된 절차에서, 참조 범위를 특정 모계 혈장 샘플 중 태아 DNA의 분획 농도 (f)에 따라 조정하는 것이 가능하다. f 값은 태아가 남성인 경우 Y 염색체에 대해 맵핑가능한 서열을 이용하여 서열분석 데이타세트로부터 측정될 수 있다. f 값은 또한 멀도 분석에서, 예를 들어 태아 추성 마커를 사용하여 (문헌 [Chan KCA et al 2006 Clin Chem 52, 2211-8]) 또는 단일 뉴클레오티드 다형성 분석으로부터 측정될 수 있다.

[0295] 한 측면에서, 표본 중의 핵산 풀을 100% 미만의 게놈 적용범위에서 서열분석하고, 포획된 핵산 분자의 비율 중에서, 대부분의 각각의 핵산 종만을 얻단 서열분석하는 경우에, 특정 유전자의 또는 염색체의 용량 불균형이 또한 정량적으로 측정될 수 있다. 다시 말하면, 유전자의 또는 염색체의 용량 불균형은 표본의 모든 맵핑가능한 서열분석된 태그 중 상기 유전자의 표출 백분율로부터 추론된다.

[0296] 대규모 병렬 게놈 서열분석 접근법에 대한 한 측면에서, 모든 염색체로부터의 대표적인 데이타는 동시에 생성될 수 있다. 특정 절편의 기원은 해당 시간 전에 선택되지 않는다. 서열분석은 랜덤하게 이루어지고, 이어서 데이타베이스 조사를 수행하여 상기 특정 절편이 어디로부터 유래되었는지 알 수 있다. 이것은 염색체 21로부터의 특정 절편 및 염색체 1로부터의 또다른 절편을 동족시키는 상황과 대비된다.

[0297] 한 예에서, 이러한 서열의 부분은 상기 설명 예에서 염색체 21과 같은 이수성에 포함되는 염색체로부터 유래할 것이다. 이러한 서열분석 실행으로부터 다른 서열이 다른 염색체로부터 유래될 것이다. 다른 염색체와 비교되는 염색체 21의 상대적 크기를 고려함으로써, 참조 범위 내에서, 상기 서열분석 실행으로부터의 염색체 21-특이적 서열의 정규화 도수를 얻을 수 있다. 태아가 삼염색체 21을 갖는 경우, 서열분석 실행으로부터의 염색체 21-유래된 서열의 정규화 도수는 증가할 것이고, 따라서 삼염색체 21이 검출된다. 정규화 도수의 변화 정도는 분석된 샘플 중 태아 핵산의 분획 농도에 따라 좌우될 것이다.

[0298] 한 실시양태에서, 본 발명자들은 인간 게놈 DNA 및 인간 혈장 DNA 샘플의 단일-말단 서열분석을 위해 일루미나 게놈 분석기를 사용하였다. 일루미나 게놈 분석기는 유동 셀로 명명된 고체 표면 상에서 포획된, 블로닝에 의해 확장된 단일 DNA 분자를 서열분석한다. 각각의 유동 셀은 8개의 개별 표본 또는 표본 풀의 서열분석을 위한 8개 레인을 포함한다. 각각의 레인은 오직 인간 게놈 서열의 30억개의 염기쌍의 분획인 약 200Mb의 서열을 생성할 수 있다. 각각의 게놈 DNA 또는 혈장 DNA 샘플은 유동 셀의 한 레인을 이용하여 서열분석하였다. 생성된 짧은 서열 태그를 인간 참조 게놈에 대해 정렬시키고, 염색체 기원을 기록하였다. 각각의 염색체에 대해 정렬된 서열분석된 개별 태그의 총 갯수를 인간 참조 게놈 또는 비질환 대표 표본으로부터 예상되는 바와 같은 각각의 염색체의 상대적인 크기와 비교하여 표로 만들었다. 이렇게 하여, 염색체 획득 또는 소실을 확인하였다.

[0299] 기재된 접근법은 현재 기재된 유전자/염색체 용량 질환의 하나의 예시일 뿐이다. 별법으로, 제어형 말단의 서열분석을 수행할 수 있다. 문헌 [Campbell et al, Nat Genet 2008; 40: 722-729]에 기재된 바와 같이 참조 게놈에서 예상되는 것으로부터 서열분석된 절편의 갯수를 비교하는 대신, 정렬된 서열분석된 태그의 갯수를 카운팅하고, 염색체 위치에 따라 분류하였다. 염색체 영역 또는 전체 염색체의 획득 또는 소실은 태그 카운트를 참

조 게놈에서의 예상되는 염색체 크기 또는 비율한 대표 표본의 예상되는 염색체 크기와 비교함으로써 측정하였다.

[0300] 또다른 실시양태에서, 실행 시 서열분석되는 핵산 풀의 분획은 서열분석 전에 추가로 하위 선택된다. 예를 들어, 올리고뉴클레오티드 어레이와 같이 혼성화를 기초로 하는 기술을 이용하여, 특정 염색체, 예를 들어 잠재적 이수체 염색체 및 시험된 이수성에 포함되지 않는 다른 염색체(들)로부터의 핵산 서열을 먼저 하위 선택할 수 있다. 또다른 예는 샘플 풀로부터의 핵산 서열의 목적 하위-집단이 서열분석 이전에 하위 선택되거나 또는 풍부화된다는 것이다. 예를 들어, 모체 혈장 중 태아 DNA 분자가 모체 배경 DNA 분자보다 더 짧은 절편으로 구성되어 있는 것으로 보고되어 있다 (문헌 [Chan et al Clin Chem 2004; 50: 88-92]). 따라서, 분자 크기에 따라 샘플 내 핵산 서열을 분획화하기 위해 당업자들에게 공지된 하나 이상의 방법, 예를 들어 겔 전기영동 또는 크기 배제 컬럼, 또는 미세유체-기반 접근법에 의한 방법을 사용할 수 있다. 그러나, 별법으로, 모체 혈장 중의 무-세포 태아 DNA 분석 예에서는, 태아 핵산 부분이 모체 배경을 억제하는 방법에 의해, 예컨대 포름알데히드의 첨가에 의해 풍부화될 수 있다 (문헌 [Dhallan et al JAMA 2004; 291 : 1114-9]).

[0301] 표준 454 플랫폼, 어플라이드 바이오시스템즈 솔리드 플랫폼, 일루미나 트루 싱글 몰레큘 DNA 서열분석 기술, 퍼시픽 바이오사이언시즈의 단일 분자, 실시간 (SMRT™) 기술 및 나노기공 서열분석에 의해서와 같은 다른 단일 분자 서열분석 전략이 본 출원에서 유사하게 사용될 수 있다.

[0302] 결과 및 (예를 들어, 서열분석 및 파라미터 계산을 위한) 추가 논의의 예는 공동 출원된 출원 "게놈 서열분석을 사용한 태아 염색체 이수성 검출" (대리인 관리 번호 016285-005220US)에서 발견될 수 있고, 이것은 본원에 참고로 포함된다. 분별값을 측정하기 위해 본원에 기재된 방법은, 예를 들어 본 섹션에 기재된 바와 같이, 반응을 서열분석하는 경우 실행될 수 있음을 주목한다.

[0303] 모체 혈장 중 태아 DNA의 분획 농도의 측정은 또한 서열분석 전략과 별도로 수행될 수 있다. 예를 들어, Y 염색체 DNA 농도는 실시간 PCR, 미세유체 PCR 또는 질량 분석기를 사용하여 따로 측정할 수 있다. 사실, 태아 DNA 농도는 여성 태아에 적용할 수 있는 Y 염색체 이외의 유전자좌를 사용하여 측정될 수 있다. 예를 들어, 챈 등(Chan et al)은 태아-유래된 메틸화된 RASSF1A 서열이 모체 유래의 비메틸화된 RASSF1A 서열을 배경으로 하여 임신 여성의 혈장에서 검출됨을 밝혀냈다 (문헌 [Chan et al, Clin Chem 2006;52:2211-81]). 따라서, 태아 DNA의 분획 농도는 메틸화된 RASSF1A 서열의 양을 전체 RASSF1A (메틸화 및 비메틸화 RASSF1A) 서열의 양으로 나눔으로써 측정될 수 있다.

[0304] 혈액 응고 등인 모체 혈액 세포로부터 DNA가 방출되기 때문에, 본 발명을 실행하기 위해서는 모체 혈장이 모체 혈청보다 바람직할 것으로 예상된다. 따라서, 혈청을 사용하는 경우, 태아 DNA의 분획 농도가 모체 혈청에서보다 모체 혈장에서 더 낮을 것으로 예상된다. 다르게 말하면, 모체 혈장을 사용하는 경우, 동일한 임신 여성으로부터 동시에 얻어지는 혈장 샘플과 비교했을 때, 진단하고자 하는 태아 염색체 이수성에 대해 더 많은 서열이 생성될 필요가 있을 것으로 예상된다.

[0305] 그러나, 태아 DNA의 분획 농도를 측정하기 위한 또다른 대안 방법은 임신 여성 및 태아 사이의 다형성 차이의 정량화를 동반하는 것일 것이다 (문헌 [Dhallan R, et al. 2007 Lancet, 369, 474-481]). 이러한 방법의 예는 임신 여성이 동형접합성이고 태아가 이형접합성인 다형성 부위를 표적으로 할 것이다. 태아-특이적 대립유전자의 양을 공통 대립유전자의 양과 비교하여 태아 DNA의 분획 농도를 측정할 수 있다.

[0306] 비교 게놈 혼성화, 마이크로어레이 비교 게놈 혼성화, 정량적 실시간 중합효소 연쇄 반응 (하나 이상의 특이적 서열(들)을 검출하고 정량화함)을 비롯한, 염색체 이상을 검출하기 위한 기존의 기술과 달리, 대규모 병렬 서열분석은 미리 측정되거나 또는 미리 한정된 세트의 DNA 서열의 검출 또는 분석에 따라 좌우되지 않는다. 표본 풀로부터의 DNA 분자의 랜덤한 대표 분획을 서열분석한다. 다양한 염색체 영역에 대해 정렬된 상이한 서열 태그의 갯수는 종양성 DNA를 함유하거나 또는 함유하지 않은 표본 사이에서 비교한다. 염색체 이상은 표본에서 임의의 소정 염색체 영역에 대해 정렬된 서열의 갯수 (또는 백분율)의 차이로 나타낼 것이다.

[0307] 또다른 예에서, 혈장 무-세포 DNA에 대한 서열분석 기술을 이용하여 특정 암의 검출을 위해 혈장 DNA에서의 염색체 이상을 검출할 수 있다. 상이한 암은 한 세트의 전형적인 염색체 이상을 갖는다. 다중 염색체 영역의 변화 (증폭 및 결실)가 사용될 수 있다. 따라서, 증폭된 영역에 대해 정렬된 서열의 비율이 증가하고, 감소된 영역에 대해 정렬된 서열의 비율이 감소될 것이다. 염색체 당 표출 백분율은 전체 게놈과 관련하여 임의의 소정 염색체의 게놈 표출의 백분율로서 표현된 참조 게놈 중 각각의 상응하는 염색체에 대한 크기와 비교할 수 있다. 직접 비교 또는 참조 염색체에 대한 비교가 또한 사용될 수 있다.

[0309] VIII. 돌연변이 검출

[0310] 모체 혈장 중의 태아 DNA는 소량의 검단으로 존재하며, 모체 혈장 DNA 중 평균 3% 내지 6%가 태아에 의한다. 이러한 이유로 인해서, 본 분야에서 행해진 기존 연구의 대부분은 태아가 부계로부터 물려받고 모체 혈장 중의 대다수의 모체 DNA 배경과 구별되는 DNA 표적을 검출하는데 집중되었다. 이와 같이 기존에 검출된 표적의 예는 Y 염색체상의 SRY 유전자 [Lo YMD et al. 1998 Am J Hum Genet, 62, 768-775] 및 모계가 RhD-음성인 경우의 RHD 유전자 [Lo YMD et al. 1998 N Engl J Med, 339, 1734-1738]를 포함한다.

[0311] 태아 돌연변이 검출을 위해서 모체 혈장을 사용하는 기존의 전략은 부계가 보인자(carrier)인 상염색체 우성 상태로 제한되며, 부계 및 모계가 상이한 돌연변이를 갖는 경우에는 적절적인 돌연변이 검출을 수행하거나 또는 연관 분석 [Ding C, et al 2004 Proc Natl Acad Sci USA 101, 10762-10767]을 실시하여 상염색체 열성 질환은 제외한다. 이러한 기존 전략에는 유의한 한계가 있었다. 예를 들어, 남성 파트너와 여성 파트너가 동일한 돌연변이를 갖는 방에서는 모체 혈장 중에서의 직접적인 돌연변이 검출로는 의미있는 태아 진단을 실시할 수 없다.

[0312] 이러한 시나리오는 도 22에 예시되어 있다. 상기 시나리오에서는, 3종의 가능한 태아 유전자형, NN, NM 및 MM이 존재하는데, 여기서 N은 정상적인 대립유전자를 나타내고, M은 돌연변이 대립유전자를 나타낸다. 돌연변이 대립유전자의 예는 낭포성 섬유증, 베타-지중해빈혈, 알파-지중해빈혈, 겸상적혈구 빈혈, 취수 관위축증, 선천성 부신 과형성 등을 일으키는 것을 포함한다. 이러한 장애의 다른 예는 OMIM (Online Mendelian Inheritance in Man) [www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM&itool=toolbar]에서 찾을 수 있다. 모체 혈장에서, 대부분의 DNA는 모계로부터 유래되며 NN일 것이다. 상기한 3종의 태아 유전자형 중 어느 것에서도, 모체 혈장 중에서의 독특한 검출을 허용하는 임의의 독특한 태아 대립유전자는 존재하지 않는다. 따라서, 통상적인 전략은 이러에 적용될 수 없다.

[0313] 본원에 기재한 실시양태는 이러한 시나리오를 취급할 수 있다. 모계 및 태아가 둘다 NM인 시나리오에서는 N 대립유전자 및 M 대립유전자가 대립유전자 균형을 이루고 있을 것이다. 그러나, 모계가 NM이고 태아가 NN인 경우에는 모체 혈장 중에서 N 대립유전자가 과다표출된 대립유전자 불균형이 일어날 것이다. 한편, 모계가 MM이고 태아가 NM인 경우에는 모체 혈장 중에서 M 대립유전자가 과다표출된 대립유전자 불균형이 일어날 것이다. 따라서, 태아 돌연변이 검출을 위한 귀무가설에서는 태아가 NM 유전자형인 경우에 대립유전자 불균형이 존재하지 않는다고 가정한다. 대립가설에서는, 대립유전자 불균형이 존재하고 N 또는 M 대립유전자 중 어느 것이 과다표출되었는지의 여부에 따라 태아 유전자형은 NN 또는 MM일 수 있다고 가정한다.

[0314] 대립유전자 불균형의 존재 또는 부재는 본원에 기재한 실시양태를 사용한 디지탈 PCR로 결정될 수 있다. 최빈 예 시나리오에서, 특정 부피의 모체 혈장은 모계로부터의 50개 세포 및 태아로부터의 50개 세포의 100개 세포로부터 방출된 DNA를 함유한다. 따라서, 이러한 부피의 혈장 중에서 태아 DNA의 분획 농도는 50%이다. 모계가 유전자형 NM인 경우에는 50개의 N 대립유전자 및 50개의 M 대립유전자가 모계에 의한다. 태아가 유전자형 NM인 경우에는 50개의 N 대립유전자 및 50개의 M 대립유전자가 태아에 의한다. 따라서, N 대립유전자와 M 대립유전자 (각각 총 100개 카피) 사이에는 대립유전자 불균형이 존재하지 않는다. 한편, 태아가 NN 유전자형인 경우에는 이러한 부피의 혈장 중에 100개의 태아-유래의 N 대립유전자가 존재할 것이다. 따라서, 총 150개의 N 대립유전자와 50개의 M 대립유전자가 존재할 것이다. 즉, N과 M 사이에는 N이 M에 대하여 3:1의 비율로 과다표출된 대립유전자 불균형이 존재할 것이다.

[0315] 반대의 상황에서, 태아가 MM 유전자형인 경우에는 이러한 부피의 혈장 중에 100개의 태아-유래의 M 대립유전자가 존재할 것이다. 따라서, 150개의 M 대립유전자와 50개의 N 대립유전자가 존재할 것이다. 즉, N과 M 사이에는 M이 N에 대하여 3:1의 비율로 과다표출된 대립유전자 불균형이 존재할 것이다. 이러한 대립유전자 불균형은 디지탈 PCR로 결정될 수 있다. 양성 웰의 수가 더 적은 대립유전자가 참조 주형으로 간주된다. 디지탈 RNA-SNP 및 디지탈 RCD 분석과 유사하게, 디지탈 PCR 실험에서 대립유전자의 실제 분포는 푸아송 확률 밀도 함수에 의해 좌우된다. 따라서, 본 시나리오에서는 대립유전자 불균형의 이론적인 정도가 3:1이거만, 대립유전자 불균형의 예상 정도는 디지탈 PCR 분석 동안에 웰 당 평균 주형 농도에 따라 달라질 것이다. 따라서, SPRT 분석에서의 같은, 웰 당 평균 참조 주형 농도 (m_c)에 적절한 해석 분열값이 사례 분류에 사용될 필요가 있다.

[0316] 추가로, 결정할 필요가 있는 대립유전자 불균형의 정도는 태아 DNA의 분획 농도에 따라 달라진다. 상기한 예와는 반대로, 특정 부피의 모체 혈장이 모계로부터의 90개 세포 및 태아로부터의 10개 세포의 100개 세포로부터 방출된 DNA를 함유하는 경우를 가정한다. 따라서, 이러한 부피의 혈장 중에서 태아 DNA의 분획 농도는 10%이

다. 모계가 유전자형 XX인 경우에는 90개의 X 대립유전자 및 90개의 X 대립유전자가 모계에 의한다. 태아가 유전자형 XX인 경우에는 10개의 X 대립유전자 및 10개의 X 대립유전자가 태아에 의한다. 따라서, X 대립유전자와 X 대립유전자 (각각 총 100개 카피) 사이에는 대립유전자 불균형이 존재하지 않는다. 한편, 태아가 XY 유전자형인 경우에는 이러한 부피의 혈장 중에 20개의 태아-유래의 X 대립유전자가 존재할 것이다. 따라서, 총 110개의 X 대립유전자와 90개의 X 대립유전자가 존재할 것이다.

[0317] 즉, X과 X 사이에는 X이 X에 대하여 과다표출된 대립유전자 불균형이 존재할 것이다. 반대의 상황에서, 태아가 XX 유전자형인 경우에는 이러한 부피의 혈장 중에 20개의 태아-유래의 X 대립유전자가 존재할 것이다. 따라서, 110개의 X 대립유전자와 90개의 X 대립유전자가 존재할 것이다. 즉, X과 X 사이에는 X이 과다표출된 대립유전자 불균형이 존재할 것이다. 태아 DNA의 분획 농도가 10%인 경우에 대립유전자 불균형의 이론적인 정도는 110:90일 것이고, 이는 상기 예에서 나타낸 바와 같이 50% 태아 DNA가 존재하는 경우의 3:1 비율과 상이하다. 따라서, SPRT 분석에서와 같은, 태아 DNA의 분획 농도에 적절한 해석 분별값이 사례 분류에 사용될 필요가 있다.

[0318] 따라서, 혈장 DNA를 추출한 것이다. 혈장 샘플 중 모체 및 태아 DNA의 양은 예를 들어 기존에 수립된 실시간 PCR 검정 [Lo, et al. 1998 Am J Hum Genet 62, 768-775] 또는 당업자에게 공지된 다른 유형의 정량화 수단, 예를 들어 SNP 마커 [Dhallan R et al., 2007 Lancet, 369, 474-481] 및 태아 후성 마커 [Chan KCA et al., 2006 Clin Chem, 52, 2211-2218]를 이용하여 정량한다. 태아 DNA의 백분율(%)을 계산한다. 이후, 디지탈 PCR 분석 동안에 각 반응 웰이 평균 1개의 주형 분자 (X 대립유전자일 수도 있고, 또는 X 대립유전자일 수도 있음)를 함유하도록 정량된 혈장 DNA 샘플 (예를 들어 희석하거나 농축된 것)을 제조한다. 환상의 프라이머, 및 2개의 대프란 프로브 (1개는 X 대립유전자에 특이적이고, 다른 하나는 X 대립유전자에 특이적임)를 사용하여 디지탈 PCR 분석을 수행한다. 오직 X에만 양성인 웰의 수 및 오직 X에만 양성인 웰의 수를 계수한다. 이들 웰의 비율을 사용하여 대립유전자 불균형의 증거가 있는지 여부를 결정한다. 대립유전자 불균형의 통계적인 증거는 당업자에게 공지된 방법, 예컨대 SPRT를 사용하여 찾을 수 있다. 이러한 분석의 한 변형법에서는 X에만 양성이거나 X 및 X에 양성인 웰의 수를 계수하는 것도 가능하고, X에만 양성이거나 X 및 X에 양성인 웰의 수를 계수하는 것도 가능하며, 이들 계수치의 비율을 도출하는 것도 가능하다. 한편 더, 대립유전자 불균형의 통계적인 증거는 당업자에게 공지된 방법, 예컨대 SPRT를 사용하여 찾을 수 있다.

[0319] 디지탈 RMD (상대적 돌연변이 양(Relative Mutant Dosage))라고 불리는 태아 유전자 돌연변이의 양 결정을 여성/남성 (XX/XY) DNA 혼합물을 사용하여 검증하였다. 남성 및 여성으로부터의 혈액 세포 DNA를 남성 DNA와 각각 혼합하여, 도 24a에 나타낸 바와 같이, XY의 배경에 XX 또는 XY 유전자형을 각각 25% 및 50%의 분획 농도로 갖는 샘플을 수득한다.

[0320] 추가로, 혈액 세포 샘플을 또한 12명의 남성 및 12명의 여성 대상체로부터 수득하였다. 여성 혈액 세포 DNA (유전자형 XX)를 3배 과량의 남성 혈액 세포 DNA (유전자형 XY)와 각각 혼합하여, XY 유전자형을 갖는 DNA 75%의 배경에 XX 유전자형을 갖는 DNA 25%를 함유하는 12개의 DNA 혼합물을 생성하였고, 이 결과는 도 24b에 나타냈다.

[0321] SPRT 분석의 목적은 배경 DNA에 존재하는 소량의 유전자형을 결정하는 것이었다. 75% XY DNA의 배경에 25% XX DNA를 함유하는 DNA 혼합물에서, 소량의 대립유전자는 DNA 75%에서 유래된 Y이다. 샘플 중의 DNA 25%가 XX 유전자형이기 때문에, 샘플 중에 총 200개의 DNA 분자가 존재하는 경우에는 150개의 분자가 XY 개체에서 유래된 것이다. 따라서, Y 대립유전자의 수는 75개일 것으로 예상된다. 또한, 남성 DNA (유전자형 XY)의 비율에 의한 X 대립유전자의 수 역시 75개이다. 여성 (유전자형 XX)에 의한 X 대립유전자의 수는 50 (= 2×25)이다. 따라서, X/Y 비율은 125/75 = (1 + 25%)/(1 - 25%) = 5/3이다.

[0322] 본 연구의 두번째 부분에서, 혈액 세포 샘플을 베타-글로빈 유전자, 즉, 헤모글로빈, 베타 (HBB) 유전자에 HbE (G→A) 및 CD41/42 (CTTT/-) 돌연변이를 갖는 남성 및 여성 대상체로부터 수득하였다. 모든 가능한 유전자형 (MM, MN 또는 NN)을 갖는 남성 태아를 보유하는 이형접합 모체 DNA (여기서, M = 돌연변이, N = 야생형))로부터 수득한 모체 혈장 샘플을 모방하기 위해서, 야생형 대립유전자에 대해 동형접합 (NN)이거나 2종의 돌연변이 중 하나에 대해 이형접합 (MN)인 남성으로부터의 혈액 세포 DNA를 동일 돌연변이에 대하여 이형접합 (MN)인 여성으로부터 수집한 혈액 세포 DNA 샘플과 각각 혼합하였다. 이로써, 다양한 분획 농도의 남성/돌연변이 DNA를 갖는 DNA 혼합물이 생성되었다. 또한, CD41/42 결실에 대해 동형접합 (MM)인 여성으로부터의 혈액 세포 DNA 샘플도 사용하여 DNA 혼합물을 제조하였다. SPRT 분류를 위한 남성 비율을 정확하게 하기 위해서, 각 DNA 혼합물 중 남성 DNA의 분획 농도를 ZFY/X 결정으로 결정하였다.

[0323] 디지털 ZFY/X 검정을 이용하여 SRY를 검증하고 또한 DNA 혼합물 중 남성 DNA의 분획 농도를 결정하였다. 염색체 X에서의 아연 핑거(Zinc Finger) 단백질 서열 (ZFX) 및 염색체 Y에서의 아연 핑거 단백질 서열 (ZFY)의 앰플리콘을 디지털 PCR 분석으로 검정하였다. 전방향 프라이머 5'-CAAGTGCTGGACTCAGATGTAACTG-3' (서열 47) 및 역방향 프라이머 5'-TGAAGTAATGTCAGAAGCTAAAACATCA-3' (서열 48)를 사용하여 ZFX 및 ZFY 유전자좌의 87-bp 앰플리콘을 우선 동시 증폭시켰다. 염색체 X 및 Y 좌라로그를 구별하는 2종의 염색체-특이적 테크만 프로브를 설계하였고, 이의 서열은 각각 5'-(VIC)TCTTTAGCACATTGCA(MGBNFQ)-3' (서열 49) 및 5'-(FAM)TCTTTACCACACTGCAG(MGBNFQ)-3' (서열 50)였다.

[0324] DNA 혼합물 중의 돌연변이 양을 돌연변이 대립유전자에 대한 정상 대립유전자의 디지털 PCR 분석으로 결정하였다. BbE 돌연변이의 경우, 정상 및 돌연변이 대립유전자의 87-bp 앰플리콘을 전방향 프라이머 5'-GGGCAAGGTGAACGTGGAT-3' (서열 51) 및 역방향 프라이머 5'-CTATTGGTCTCCTTAAACCTGTCTTGTAA-3' (서열 52)로 우선 동시 증폭시켰다. 정상 (G) 및 돌연변이 (A) 대립유전자 사이를 구별하는 2종의 대립유전자-특이적 테크만 프로브를 설계하였고, 이의 서열은 각각 5'-(VIC) TTGGTGGTGAGGCC (MGBNFQ)-3' (서열 53) 및 5'-(FAM) TTGGTGGTAAGGCC (MGBNFQ)-3' (서열 54)였다. BbE 돌연변이에 대한 결과를 도 25에 나타냈다.

[0325] CD41/42 결실 돌연변이의 경우, 정상 및 돌연변이 대립유전자의 87-bp 및 83-bp 앰플리콘을 전방향 프라이머 5'-TTTTCCCACCCTTAGCCTGC-3' (서열 55) 및 역방향 프라이머 5'-ACAGCATCAGGAGTGGACAGATC-3' (서열 56) 각각을 사용하여 우선 동시 증폭시켰다. 정상 (결실 없음) 및 돌연변이 (결실 있음) 대립유전자 사이를 구별하는 2종의 대립유전자-특이적 테크만 프로브를 설계하였고, 이의 서열은 각각 5'-(VIC) CAGAGGTTCTTTGAGTCCT (MGBNFQ)-3' (서열 57) 및 5'-(FAM) AGAGGTTGAGTCCTT (MGBNFQ)-3' (서열 58)였다. BbE 돌연변이에 대한 결과를 도 26a 및 도 26b에 나타냈다.

[0326] 이러한 실험은 12,765 디지털 어레이즈 (플루이다임)를 사용하는 바이오마크™ 시스템 (플루이다임)에서 수행하였다. 1개의 패널에 대한 반응은 2× 테크만 유니버설 PCR 마스타 믹스 (어플라이드 바이오시스템즈)를 사용하여 반응 부피 10 ㎕로 설정하였다. CD41/42 및 ZFY/X 검정의 경우, 각 반응물은 1× 테크만 유니버설 PCR 마스타 믹스, 900 nM씩의 각 프라이머, 125 nM씩의 각 프로브 및 1 ng/㎕의 DNA 혼합물 3.5 ㎕를 함유하였다. BbE 검정의 경우, 정상 (G) 및 돌연변이 (A) 대립유전자를 표적으로 하는, 각각 250 nM 및 125 nM의 프로브를 평가하였다. 샘플/검정 혼합물을 나노플렉스™ IFC 제어기 (플루이다임)로 디지털 어레이에 로딩하였다. 반응은 신호 검출을 위해 바이오마크™ 시스템에서 수행하였다. 반응을 50℃에서 2분 동안 개시한 후에 95℃에서 10분 동안 진행하고, 95℃, 15초 및 57℃ (ZFY/X 및 CD41/42의 경우) 또는 56℃ (BbE의 경우), 1분의 50회 주기를 수행하였다. 각 경우마다 적어도 1개의 반응 패널을 사용하였고, 결정이 이루어질 수 있을 때까지 미분류 상태로 있는 샘플을 위한 여분의 패널로부터 데이타를 합하였다.

[0327] 또한, 디지털 PCR이 당업자에게 공지된 방법, 예를 들어 미세유체 칩, 나노리터 PCR 마이크로플레이트 시스템, 에멀젼 PCR, 콜로니 PCR, 롤링-써클 증폭법, 프라이머 연장법 및 질량 분광법을 이용하여 수행될 수 있다는 것이 당업자에게 명백할 것이다.

[0329] IX. 양을 이용한 예

[0330] 한 실시양태에서, 본 방법을 수행하여 암상 증양에서 발생할 수 있는 것과 같이 대립유전자 비율이 벗어나거나 벗어나서 않은 샘플을 분류할 수 있다. 한 측면에서, 각 경우에 오직 A 대립유전자에 대해서만 양성 신호를 갖는 웰의 수, 오직 G 대립유전자에 대해서만 양성 신호를 갖는 웰의 수, 및 2종의 대립유전자 둘다에 대해서 양성 신호를 갖는 웰의 수를 디지털 PCR로 결정하였다. 참조 대립유전자는 양성 웰의 수가 더 적은 대립유전자로 정의된다 (2종의 대립유전자가 둘다 동일한 수의 양성 웰을 갖는 시나리오가 있다고 한다면, 둘 중 어느 것이라도 참조 대립유전자로 사용될 수 있음). 참조 대립유전자에 대해 음성인 웰의 총 수를 사용하여, 다른 대립유전자가 양성인지 여부와는 관계 없이 푸아송 확률 밀도 함수에 따라 웰 당 참조 대립유전자의 평균 농도 추론치 (m_c)를 계산하였다. 본 연구진은 상기 계산을 예시하는 가정적인 예를 사용한다.

[0331] 96웰 반응에서, 30개 웰은 A 대립유전자 대해 양성이고, 24개 웰은 G 대립유전자에 대해 양성이며, 28개 웰은 이들 2개의 대립유전자 모두에 대해 양성이다. A 대립유전자에 대하여 양성인 웰의 수가 더 적기 때문에, A 대립유전자를 참조 대립유전자로 간주한다. 참조 대립유전자에 대해 음성인 웰의 수는 96 - 30 - 28 = 38개이다.