

[0645] Set of NR free floating DNA sequence measurements $S=(s_1, \dots, s_{NR})$. Since this method utilizes the SNP measurements, all sequence data that corresponds to non-polymorphic loci can be disregarded. In a simplified version, where we have (A,B) counts on each SNP, where A and B correspond to the two alleles present at a given locus, S can be written as $S=((a_1, b_1), \dots, (a_N, b_N))$, where a_i is the A count on SNP i, b_i is the B count on SNP i, and $\sum_{i=1:N} (a_i + b_i) = NR$

[0646] Parent data consisting of

[0647] Genotypes from a SNP microarray or other intensity based genotyping platform: mother $M=(m_1, \dots, m_N)$, father $F=(f_1, \dots, f_N)$, where $m_i, f_i \in (AA, AB, BB)$.

[0648] AND/OR sequence data measurements: NRM mother measurements $SM=(sm_1, \dots, sm_{nrm})$, NRF father measurements $SF=(sf_1, \dots, sf_{nrf})$. Similar to the above simplification, if we have (A,B) counts on each SNP $SM=((am_1, bm_1), \dots, (am_N, bm_N))$, $SF=((af_1, bf_1), \dots, (af_N, bf_N))$

[0649] Collectively, the mother, father child data are denoted as $D = (M, F, SM, SF, S)$. Note that the parent data is desired and increases the accuracy of the algorithm, but is NOT necessary, especially the father data. This means that even in the absence of mother and/or father data, it is possible to get very accurate copy number results.

[0650] It is possible to derive the best copy number estimate (H^*) by maximizing the data log likelihood $LIK(D|H)$ over all hypotheses (H) considered. In particular it is possible to determine the relative probability of each of the ploidy hypotheses using the joint distribution model and the allele counts measured on the prepared sample, and using those relative probabilities to determine the hypothesis most likely to be correct as follows:

$$H^* = \underset{H}{\operatorname{argmax}} LIK(D|H)$$

[0651] Similarly the a posteriori hypothesis likelihood given the data may be written as:

$$H^* = \underset{H}{\operatorname{argmax}} LIK(D|H) * \operatorname{priorprob}(H)$$

[0652] Where $\operatorname{priorprob}(H)$ is the prior probability assigned to each hypothesis H, based on model design and prior knowledge.

[0653] It is also possible to use priors to find the maximum a posteriori estimate:

$$[0654] H_{MA} = \underset{H}{\operatorname{argmax}} LIK(D|H)$$

[0655] In an embodiment, the copy number hypotheses that may be considered are:

[0656] Monosomy:

[0657] maternal H10 (one copy from mother)

[0658] paternal H01 (one copy from father)

[0659] Disomy: H11 (one copy each mother and father)

[0660] Simple trisomy, no crossovers considered:

[0661] Maternal: H21_matched (two identical copies from mother, one copy from father),
H21_unmatched (BOTH copies from mother, one copy from father)

[0662] Paternal: H12_matched (one copy from mother , two identical copies from father),
H12_unmatched (one copy from mother , both copies from father)

[0663] Composite trisomy, allowing for crossovers (using a joint distribution model):

[0664] maternal H21 (two copies from mother, one from father),

[0665] paternal H12 (one copy from mother, two copies from father)

[0666] In other embodiments, other ploidy states, such as nullsomy (H00), uniparental disomy (H20 and H02), and tetrasomy (H04, H13, H22, H31 and H40), may be considered.

[0667] If there are no crossovers, each trisomy, whether the origin was mitosis, meiosis I, or meiosis II, would be one of the matched or unmatched trisomies. Due to crossovers, true trisomy is usually a combination of the two. First, a method to derive hypothesis likelihoods for simple hypotheses is described. Then a method to derive hypothesis likelihoods for composite hypotheses is described, combining individual SNP likelihood with crossovers.

[0668] *LIK(D|H) for a Simple Hypothesis*

[0669] In an embodiment, LIK(D|H) may be determined for simple hypotheses, as follows. For simple hypotheses H, LIK(H), the log likelihood of hypothesis H on a whole chromosome, may be calculated as the sum of log likelihoods of individual SNPs, assuming known or derived child fraction cf. In an embodiment it is possible to derive cf from the data.

[0670] $LIK(D|H) = \sum_i LIK(D|H, cf, i)$

[0671] This hypothesis does not assume any linkage between SNPs, and therefore does not utilize a joint distribution model.

[0672] In some embodiments, the Log Likelihood may be determined on a per SNP basis. On a particular SNP i, assuming fetal ploidy hypothesis H and percent fetal DNA cf, log likelihood of observed data D is defined as:

[0673] $LIK(D|H, i) = \log P(D|H, cf, i) =$

$$\log(\sum_{m,f,c} P(D|m, f, c, H, cf, i)P(c|m, f, H)P(m|i)P(f|i))$$

[0674] where m are possible true mother genotypes, f are possible true father genotypes, where $m, f \in \{AA, AB, BB\}$, and c are possible child genotypes given the hypothesis H. In particular, for monosomy $c \in \{A, B\}$, for disomy $c \in \{AA, AB, BB\}$, for trisomy $c \in \{AAA, AAB, ABB, BBB\}$.

[0675] Genotype prior frequency: $p(m|i)$ is the general prior probability of mother genotype m on SNP i, based on the known population frequency at SNP I, denoted pA_i . In particular

$$[0676] p(AA|pA_i) = (pA_i)^2, p(AB|pA_i) = 2(pA_i) * (1 - pA_i), p(BB|pA_i) = (1 - pA_i)^2$$

[0677] Father genotype probability, $p(f|i)$, may be determined in an analogous fashion.

[0678] True child probability: $p(c|m, f, H)$ is the probability of getting true child genotype = c, given parents m, f, and assuming hypothesis H, which can be easily calculated. For example, for H11, H21 matched and H21 unmatched, $p(c|m, f, H)$ is given below.

		p(c m,f,H)											
m	f	H11			H21 matched				H21 unmatched				
		AA	AB	BB	AAA	AAB	ABB	BBB	AAA	AAB	ABB	BBB	
AA	AA	1	0	0	1	0	0	0	1	0	0	0	
AB	AA	0.5	0.5	0	0.5	0	0.5	0	0	1	0	0	
BB	AA	0	1	0	0	0	1	0	0	0	1	0	
AA	AB	0.5	0.5	0	0.5	0.5	0	0	0.5	0.5	0	0	
AB	AB	0.25	0.5	0.25	0.25	0.25	0.25	0.25	0	0.5	0.5	0	
BB	AB	0	0.5	0.5	0	0	0.5	0.5	0	0	0.5	0.5	
AA	BB	0	1	0	0	1	0	0	0	1	0	0	
AB	BB	0	0.5	0.5	0	0.5	0	0.5	0	0	1	0	
BB	BB	0	0	1	0	0	0	1	0	0	0	1	

[0679]

[0680] Data likelihood: $P(D|m, f, c, H, i, cf)$ is the probability of given data D on SNP i, given true mother genotype m, true father genotype f, true child genotype c, hypothesis H and child fraction cf. It can be broken down into the probability of mother, father and child data as follows:

$$[0681] P(D|m, f, c, H, cf, i) = P(SM|m, i)P(M|m, i)P(SF|f, i)P(F|f, i)P(S|m, c, H, cf, i)$$

[0682] Mother SNP array data likelihood: Probability of mother SNP array genotype data m_i at SNP i compared to true genotype m, assuming SNP array genotypes are correct, is simply

$$P(M|m, i) = \begin{cases} 1 & m_i = m \\ 0 & m_i \neq m \end{cases}$$

[0683] Mother sequence data likelihood: the probability of the mother sequence data at SNP i, in the case of counts $S_i=(a_{m_i}, b_{m_i})$, with no extra noise or bias involved, is the binomial

probability defined as $P(SM|m,i)=P_{X|m}(a_i)$ where $X|m \sim \text{Binom}(p_m(A), a_i+b_i)$ with $p_m(A)$ defined as

m	AA	AB	BB	A	B	nocall
[0684] p(A)	1	0.5	0	1	0	0.5

[0685] Father data likelihood: a similar equation applies for father data likelihood.

[0686] Note that it is possible to determine the child genotype without the parent data, especially father data. For example if no father genotype data F is available, one may just use $P(F|f,i) = 1$. If no father sequence data SF is available, one may just use $P(SF|f,i)=1$.

[0687] In some embodiments, the method involves building a joint distribution model for the expected allele counts at a plurality of polymorphic loci on the chromosome for each ploidy hypothesis; one method to accomplish such an end is described here. Free fetal DNA data likelihood: $P(S|m, c, H, cf, i)$ is the probability of free fetal DNA sequence data on SNP i , given true mother genotype m , true child genotype c , child copy number hypothesis H , and assuming child fraction cf . It is in fact the probability of sequence data S on SNP i , given the true probability of A content on SNP i $\mu(m, c, cf, H)$

[0688] $P(S|m, c, H, cf, i) = P(S|\mu(m, c, cf, H), i)$

[0689] For counts, where $S_i=(a_i,b_i)$, with no extra noise or bias in data involved,

$$P(S|\mu(m, c, cf, H), i) = P_x(a_i)$$

[0690] where $X \sim \text{Binom}(p(A), a_i+b_i)$ with $p(A)=\mu(m, c, cf, H)$. In a more complex case where the exact alignment and (A,B) counts per SNP are not known, $P(S|\mu(m, c, cf, H), i)$ is a combination of integrated binomials.

[0691] True A content probability: $\mu(m, c, cf, H)$, the true probability of A content on SNP i in this mother/child mixture, assuming that true mother genotype = m , true child genotype = c , and overall child fraction = cf , is defined as

$$[0692] \mu(m, c, cf, H) = \frac{\#A(m)*(1-cf)+\#A(c)*cf}{n_m*(1-cf)+n_c*cf}$$

[0693] where $\#A(g)$ = number of A 's in genotype g , $n_m = 2$ is somy of mother and n_c is ploidy of the child under hypothesis H (1 for monosomy, 2 for disomy, 3 for trisomy).

[0694] *Using A Joint Distribution Model: LIK(D|H) for a Composite Hypothesis*

[0695] In some embodiments, the method involves building a joint distribution model for the expected allele counts at the plurality of polymorphic loci on the chromosome for each ploidy

hypothesis; one method to accomplish such an end is described here. In many cases, trisomy is usually not purely matched or unmatched, due to crossovers, so in this section results for composite hypotheses H21 (maternal trisomy) and H12 (paternal trisomy) are derived, which combine matched and unmatched trisomy, accounting for possible crossovers.

[0696] In the case of trisomy, if there were no crossovers, trisomy would be simply matched or unmatched trisomy. Matched trisomy is where child inherits two copies of the identical chromosome segment from one parent. Unmatched trisomy is where child inherits one copy of each homologous chromosome segment from the parent. Due to crossovers, some segments of a chromosome may have matched trisomy, and other parts may have unmatched trisomy. Described in this section is how to build a joint distribution model for the heterozygosity rates for a set of alleles; that is, for the expected allele counts at a number of loci for one or more hypotheses.

[0697] Suppose that on SNP i , $LIK(D|H_m, i)$ is the fit for matched hypothesis H_m , and $LIK(D|H_u, i)$ is the fit for unmatched hypothesis H_u , and $pc(i)$ = probability of crossover between SNPs $i-1$ and i . One may then calculate the full likelihood as:

[0698] $LIK(D|H) = \sum_E LIK(D|E, 1:N)$

[0699] where $LIK(D|E, 1:N)$ is the likelihood of ending in hypothesis E , for SNPs $1:N$. E = hypothesis of the last SNP, $E \in (H_m, H_u)$. Recursively, one may calculate:

[0700] $LIK(D|E, 1:i) = LIK(D|E, i) + \log(\exp(LIK(D|E, 1:i-1)) * (1 - pc(i)) + \exp(LIK(D|\sim E, 1:i-1)) * pc(i))$

[0701] where $\sim E$ is the hypothesis other than E (not E), where hypotheses considered are H_m and H_u . In particular, one may calculate the likelihood of $1:i$ SNPs, based on likelihood of 1 to $(i-1)$ SNPs with either the same hypothesis and no crossover, or the opposite hypothesis and a crossover, multiplied by the likelihood of the SNP i

[0702] For SNP 1, $i=1$, $LIK(D|E, 1:1) = LIK(D|E, 1)$.

[0703] For SNP 2, $i=2$, $LIK(D|E, 1:2) = LIK(D|E, 2) + \log(\exp(LIK(D|E, 1)) * (1 - pc(2)) + \exp(LIK(D|\sim E, 1)) * pc(2))$,

[0704] and so on for $i=3:N$.

[0705] In some embodiments, the child fraction may be determined. The child fraction may refer to the proportion of sequences in a mixture of DNA that originate from the child. In the context of non-invasive prenatal diagnosis, the child fraction may refer to the proportion of sequences

in the maternal plasma that originate from the fetus or the portion of the placenta with fetal genotype. It may refer to the child fraction in a sample of DNA that has been prepared from the maternal plasma, and may be enriched in fetal DNA. One purpose of determining the child fraction in a sample of DNA is for use in an algorithm that can make ploidy calls on the fetus, therefore, the child fraction could refer to whatever sample of DNA was analyzed by sequencing for the purpose of non-invasive prenatal diagnosis.

[0706] Some of the algorithms presented in this disclosure that are part of a method of non-invasive prenatal aneuploidy diagnosis assume a known child fraction, which may not always be the case. In an embodiment, it is possible to find the most likely child fraction by maximizing the likelihood for disomy on selected chromosomes, with or without the presence of the parental data

[0707] In particular, suppose that $\text{LIK}(D|H11, cf, chr) = \log \text{likelihood}$ as described above, for the disomy hypothesis, and for child fraction cf on chromosome chr . For selected chromosomes in $Cset$ (usually 1:16), assumed to be euploid, the full likelihood is:

$$[0705] \text{LIK}(cf) = \sum_{chr \in Cset} \text{LIK}(D|H11, cf, chr)$$

[0708] The most likely child fraction (cf^*) is derived as $cf^* = \underset{cf}{\operatorname{argmax}} \text{LIK}(cf)$.

[0709] It is possible to use any set of chromosomes. It is also possible to derive child fraction without assuming euploidy on the reference chromosomes. Using this method it is possible to determine the child fraction for any of the following situations: (1) one has array data on the parents and shotgun sequencing data on the maternal plasma; (2) one has array data on the parents and targeted sequencing data on the maternal plasma; (3) one has targeted sequencing data on both the parents and maternal plasma; (4) one has targeted sequencing data on both the mother and the maternal plasma fraction; (5) one has targeted sequencing data on the maternal plasma fraction; (6) other combinations of parental and child fraction measurements.

[0710] In some embodiments the informatics method may incorporate data dropouts; this may result in ploidy determinations of higher accuracy. Elsewhere in this disclosure it has been assumed that the probability of getting an A is a direct function of the true mother genotype, the true child genotype, the fraction of the child in the mixture, and the child copy number. It is also possible that mother or child alleles can drop out, for example instead of measuring true child AB in the mixture, it may be the case that only sequences mapping to allele A are measured. One may denote the parent dropout rate for genomic Illumina data d_{pg} , parent dropout rate for

sequence data d_{ps} and child dropout rate for sequence data d_{cs} . In some embodiments, the mother dropout rate may be assumed to be zero, and child dropout rates are relatively low; in this case, the results are not severely affected by dropouts. In some embodiments the possibility of allele dropouts may be sufficiently large that they result in a significant effect of the predicted ploidy call. For such a case, allele dropouts have been incorporated into the algorithm here:

[0711] Parent SNP array data dropouts: For mother genomic data M , suppose that the genotype after the dropout is m_d , then

[0712] $P(M|m, i) = \sum_{m_d} P(M|m_d, i)P(m_d|m)$

[0713] where $P(M|m_d, i) = \begin{cases} 1 & m_i = m_d \\ 0 & m_i \neq m_d \end{cases}$ as before, and $P(m_d|m)$ is the likelihood of genotype m_d after the possible dropout given the true genotype m , defined as below, for dropout rate d

m	md					
	AA	AB	BB	A	B	nocall
AA	$(1-d)^2$	0	0	$2d(1-d)$	0	d^2
AB	0	$(1-d)^2$	0	$d(1-d)$	$d(1-d)$	d^2
BB	0	0	$(1-d)^2$	0	$2d(1-d)$	d^2

[0715] A similar equation applies for father SNP array data.

[0716] Parent sequence data dropouts: For mother sequence data SM

[0717] $P(SM|m, i) = \sum_{m_d} P_{X|m_d}(am_i)P(m_d|m)$

[0718] where $P(m_d|m)$ is defined as in previous section and $P_{X|m_d}(am_i)$ probability from a binomial distribution is defined as before in the parent data likelihood section. A similar equation applies to the paternal sequence data.

[0719] Free floating DNA sequence data dropout:

[0720] $P(S|m, c, H, cf, i) = \sum_{m_d, c_d} P(S|\mu(m_d, c_d, cf, H), i)P(m_d|m)P(c_d|c)$

[0721] where $P(S|\mu(m_d, c_d, cf, H), i)$ is as defined in the section on free floating data likelihood.

[0722] In an embodiment, $p(m_d|m)$ is the probability of observed mother genotype m_d , given true mother genotype m , assuming dropout rate d_{ps} , and $p(c_d|c)$ is the probability of observed child genotype c_d , given true child genotype c , assuming dropout rate d_{cs} . If nA_T = number of A alleles in true genotype c , nA_D = number of A alleles in observed genotype c_d , where $nA_T \geq$

n_{A_D} , and similarly n_{B_T} = number of B alleles in true genotype c, n_{B_D} = number of B alleles in observed genotype c_d , where $n_{B_T} \geq n_{B_D}$ and d = dropout rate, then

$$[0723] \ p(c_d|c) = \binom{n_{A_T}}{n_{A_D}} * d^{n_{A_T}-n_{A_D}} * (1-d)^{n_{A_D}} * \binom{n_{B_T}}{n_{B_D}} * d^{n_{B_T}-n_{B_D}} * (1-d)^{n_{B_D}}$$

[0724] In an embodiment, the informatics method may incorporate random and consistent bias. In an ideal word there is no per SNP consistent sampling bias or random noise (in addition to the binomial distribution variation) in the number of sequence counts. In particular, on SNP i, for mother genotype m, true child genotype c and child fraction cf, and X = the number of A's in the set of (A+B) reads on SNP i, X acts like a $X \sim \text{Binomial}(p, A+B)$, where $p = \mu(m, c, cf, H)$ = true probability of A content.

[0725] In an embodiment, the informatics method may incorporate random bias. As is often the case, suppose that there is a bias in the measurements, so that the probability of getting an A on this SNP is equal to q, which is a bit different than p as defined above. How much different p is from q depends on the accuracy of the measurement process and number of other factors and can be quantified by standard deviations of q away from p. In an embodiment, it is possible to model q as having a beta distribution, with parameters α, β depending on the mean of that distribution being centered at p, and some specified standard deviation s. In particular, this gives $X|q \sim \text{Bin}(q, D_i)$, where $q \sim \text{Beta}(\alpha, \beta)$. If we let $E(q) = p$, $V(q) = s^2$, and parameters α, β can be derived as $\alpha = pN$, $\beta = (1-p)N$, where $N = \frac{p(1-p)}{s^2} - 1$.

[0726] This is the definition of a beta-binomial distribution, where one is sampling from a binomial distribution with variable parameter q, where q follows a beta distribution with mean p. So, in a setup with no bias, on SNP i, the parent sequence data (SM) probability assuming true mother genotype (m), given mother sequence A count on SNP i (am_i) and mother sequence B count on SNP i (bm_i) may be calculated as:

$$[0724] \ P(\text{SM}|m,i) = P_{X|m}(am_i) \text{ where } X|m \sim \text{Binom}(p_m(A), am_i+bm_i)$$

[0727] Now, including random bias with standard deviation s, this becomes:

$$[0725] \ X|m \sim \text{BetaBinom}(p_m(A), am_i+bm_i, s)$$

[0728] In the case with no bias, the maternal plasma DNA sequence data (S) probability assuming true mother genotype (m), true child genotype (c), child fraction (cf), assuming child

hypothesis H, given free floating DNA sequence A count on SNP i (a_i) and free floating sequence B count on SNP i (b_i) may be calculated as

$$[0726] \quad P(S|m, c, cf, H, i) = P_x(a_i)$$

[0729] where $X \sim \text{Binom}(p(A), a_i + b_i)$ with $p(A) = \mu(m, c, cf, H)$.

[0730] In an embodiment, including random bias with standard deviation s , this becomes $X \sim \text{BetaBinom}(p(A), a_i + b_i, s)$, where the amount of extra variation is specified by the deviation parameter s , or equivalently N . The smaller the value of s (or the larger the value of N) the closer this distribution is to the regular binomial distribution. It is possible to estimate the amount of bias, *i.e.* estimate N above, from unambiguous contexts AA|AA, BB|BB, AA|BB, BB|AA and use estimated $\hat{N}$ in the above probability. Depending on the behavior of the data, N may be made to be a constant irrespective of the depth of read $a_i + b_i$, or a function of $a_i + b_i$, making bias smaller for larger depths of read.

[0731] In an embodiment, the informatics method may incorporate consistent per-SNP bias. Due to artifacts of the sequencing process, some SNPs may have consistently lower or higher counts irrespective of the true amount of A content. Suppose that SNP i consistently adds a bias of w_i percent to the number of A counts. In some embodiments, this bias can be estimated from the set of training data derived under same conditions, and added back in to the parent sequence data estimate as:

$$[0730] \quad P(SM|m, i) = P_{X|m}(a_i) \text{ where } X|m \sim \text{BetaBinom}(p_m(A) + w_i, a_i + b_i, s)$$

[0731] and with the free floating DNA sequence data probability estimate as:

$$[0732] \quad P(S|m, c, cf, H, i) = P_x(a_i) \text{ where } X \sim \text{BetaBinom}(p(A) + w_i, a_i + b_i, s),$$

[0732] In some embodiments, the method may be written to specifically take into account additional noise, differential sample quality, differential SNP quality, and random sampling bias. An example of this is given here. This method has been shown to be particularly useful in the context of data generated using the massively multiplexed mini-PCR protocol, and was used in Examples 7 through 13. The method involves several steps that each introduce different kind of noise and/or bias to the final model:

[0733] Suppose the first sample that comprises a mixture of maternal and fetal DNA contains an original amount of DNA of size $= N_0$ molecules, usually in the range 1,000-40,000, where p = true %refs

[0734] In the amplification using the universal ligation adaptors, assume that N_1 molecules are sampled; usually $N_1 \sim N_0/2$ molecules and random sampling bias is introduced due to sampling.

The amplified sample may contain a number of molecules N_2 where $N_2 \gg N_1$. Let X_1 represent the amount of reference loci (on per SNP basis) out of N_1 sampled molecules, with a variation in $p_1 = X_1/N_1$ that introduces random sampling bias throughout the rest of protocol. This sampling bias is included in the model by using a Beta-Binomial (BB) distribution instead of using a simple Binomial distribution model. Parameter N of the Beta-Binomial distribution may be estimated later on per sample basis from training data after adjusting for leakage and amplification bias, on SNPs with $0 < p < 1$. Leakage is the tendency for a SNP to be read incorrectly.

[0735] The amplification step will amplify any allelic bias, thus amplification bias introduced due to possible uneven amplification. Suppose that one allele at a locus is amplified f times another allele at that locus is amplified g times, where $f = ge^b$, where $b=0$ indicates no bias. The bias parameter, b , is centered at 0, and indicates how much more or less the A allele get amplified as opposed to the B allele on a particular SNP. The parameter b may differ from SNP to SNP. Bias parameter b may be estimated on per SNP basis, for example from training data.

[0736] The sequencing step involves sequencing a sample of amplified molecules. In this step there may be leakage, where leakage is the situation where a SNP is read incorrectly. Leakage may result from any number of problems, and may result in a SNP being read not as the correct allele A, but as another allele B found at that locus or as an allele C or D not typically found at that locus. Suppose the sequencing measures the sequence data of a number of DNA molecules from an amplified sample of size N_3 , where $N_3 < N_2$. In some embodiments, N_3 may be in the range of 20,000 to 100,000; 100,000 to 500,000; 500,000 to 4,000,000; 4,000,000 to 20,000,000; or 20,000,000 to 100,000,000. Each molecule sampled has a probability p_g of being read correctly, in which case it will show up correctly as allele A. The sample will be incorrectly read as an allele unrelated to the original molecule with probability $1-p_g$, and will look like allele A with probability p_r , allele B with probability p_m or allele C or allele D with probability p_o , where $p_r + p_m + p_o = 1$. Parameters p_g , p_r , p_m , p_o are estimated on per SNP basis from the training data.

[0737] Different protocols may involve similar steps with variations in the molecular biology steps resulting in different amounts of random sampling, different levels of amplification and different leakage bias. The following model may be equally well applied to each of these cases. The model for the amount of DNA sampled, on per SNP basis, is given by:

$$X_3 \sim \text{BetaBinomial}(L(F(p,b), p_r, p_g), N * H(p,b))$$

[0738] where p = the true amount of reference DNA, b = per SNP bias, and as described above, p_g is the probability of a correct read, p_r is the probability of read being read incorrectly but serendipitously looking like the correct allele, in case of a bad read, as described above, and:

$$F(p,b) = pe^b/(pe^b+(1-p)), H(p,b) = (e^bp+(1-p))^2/e^b, L(p,p_r,p_g) = p*p_g + p_r*(1-p_g).$$

[0739] In some embodiments, the method uses a Beta-Binomial distribution instead of a simple binomial distribution; this takes care of the random sampling bias. Parameter N of the Beta-Binomial distribution is estimated on per sample basis on an as needed basis. Using bias correction $F(p,b)$, $H(p,b)$, instead of just p , takes care of the amplification bias. Parameter b of the bias is estimated on per SNP basis from training data ahead of time.

[0740] In some embodiments the method uses leakage correction $L(p,p_r,p_g)$, instead of just p ; this takes care of the leakage bias, *i.e.* varying SNP and sample quality. In some embodiments, parameters p_g , p_r , p_o are estimated on per SNP basis from the training data ahead of time. In some embodiments, the parameters p_g , p_r , p_o may be updated with the current sample on the go, to account for varying sample quality.

[0741] The model described herein is quite general and can account for both differential sample quality and differential SNP quality. Different samples and SNPs are treated differently, as exemplified by the fact that some embodiments use Beta-Binomial distributions whose mean and variance are a function of the original amount of DNA, as well as sample and SNP quality.

[0742] Platform modeling

[0743] Consider a single SNP where the expected allele ratio present in the plasma is r (based on the maternal and fetal genotypes). The expected allele ratio is defined as the expected fraction of A alleles in the combined maternal and fetal DNA. For maternal genotype g_m and child genotype g_c , the expected allele ratio is given by equation 1, assuming that the genotypes are represented as allele ratios as well.

$$[0743] \quad r = fg_c + (1 - f)g_m \quad (1)$$

[0744] The observation at the SNP consists of the number of mapped reads with each allele present, n_a and n_b , which sum to the depth of read d . Assume that thresholds have already been applied to the mapping probabilities and phred scores such that the mappings and allele observations can be considered correct. A phred score is a numerical measure that relates to the probability that a particular measurement at a particular base is wrong. In an embodiment, where

the base has been measured by sequencing, the phred score may be calculated from the ratio of the dye intensity corresponding to the called base to the dye intensity of the other bases. The simplest model for the observation likelihood is a binomial distribution which assumes that each of the d reads is drawn independently from a large pool that has allele ratio r . Equation 2 describes this model.

$$P(n_a, n_b | r) = p_{\text{bino}}(n_a; n_a + n_b, r) = \binom{n_a + n_b}{n_a} r^{n_a} (1 - r)^{n_b} \quad (2)$$

[0745] The binomial model can be extended in a number of ways. When the maternal and fetal genotypes are either all A or all B, the expected allele ratio in plasma will be 0 or 1, and the binomial probability will not be well-defined. In practice, unexpected alleles are sometimes observed in practice. In an embodiment, it is possible to use a corrected allele ratio $\hat{r} = 1/(n_a + n_b)$ to allow a small number of the unexpected allele. In an embodiment, it is possible to use training data to model the rate of the unexpected allele appearing on each SNP, and use this model to correct the expected allele ratio. When the expected allele ratio is not 0 or 1, the observed allele ratio may not converge with a sufficiently high depth of read to the expected allele ratio due to amplification bias or other phenomena. The allele ratio can then be modeled as a beta distribution centered at the expected allele ratio, leading to a beta-binomial distribution for $P(n_a, n_b | r)$ which has higher variance than the binomial.

[0746] The platform model for the response at a single SNP will be defined as $F(a, b, g_c, g_m, f)$ (3), or the probability of observing $n_a = a$ and $n_b = b$ given the maternal and fetal genotypes, which also depends on the fetal fraction through equation 1. The functional form of F may be a binomial distribution, beta-binomial distribution, or similar functions as discussed above.

$$F(a, b, g_c, g_m, f) = P(n_a = a, n_b = b | g_c, g_m, f) = P(n_a = a, n_b = b | r(g_c, g_m, f)) \quad (3)$$

[0747] In an embodiment, the child fraction may be determined as follows. A maximum likelihood estimate of the fetal fraction f for a prenatal test may be derived without the use of paternal information. This may be relevant where the paternal genetic data is not available, for example where the father of record is not actually the genetic father of the fetus. The fetal fraction is estimated from the set of SNPs where the maternal genotype is 0 or 1, resulting in a set of only two possible fetal genotypes. Define S_0 as the set of SNPs with maternal genotype 0 and S_1 as the set of SNPs with maternal genotype 1. The possible fetal genotypes on S_0 are 0 and

0.5, resulting in a set of possible allele ratios $R_0(f) = \{0, f/2\}$. Similarly, $R_1(f) = \{1-f/2, 1\}$. This method can be trivially extended to include SNPs where maternal genotype is 0.5, but these SNPs will be less informative due to the larger set of possible allele ratios.

[0748] Define N_{a0} and N_{b0} as the vectors formed by n_{as} and n_{bs} for SNPs s in S_0 , and N_{a1} and N_{b1} similarly for S_1 . The maximum likelihood estimate $\hat{f}$ of f is defined by equation 4.

$$\hat{f} = \arg \max_f P(N_{a0}, N_{b0}|f) P(N_{a1}, N_{b1}|f) \quad (4)$$

[0749] Assuming that the allele counts at each SNP are independent conditioned on the SNP's plasma allele ratio, the probabilities can be expressed as products over the SNPs in each set (5).

$$P(N_{a0}, N_{b0}|f) = \prod_{s \in S_0} P(n_{as}, n_{bs}|f) \quad (5)$$

$$P(N_{a1}, N_{b1}|f) = \prod_{s \in S_1} P(n_{as}, n_{bs}|f)$$

[0750] The dependence on f is through the sets of possible allele ratios $R_0(f)$ and $R_1(f)$. The SNP probability $P(n_{as}, n_{bs}|f)$ can be approximated by assuming the maximum likelihood genotype conditioned on f . At reasonably high fetal fraction and depth of read, the selection of the maximum likelihood genotype will be high confidence. For example, at fetal fraction of 10 percent and depth of read of 1000, consider a SNP where the mother has genotype zero. The expected allele ratios are 0 and 5 percent, which will be easily distinguishable at sufficiently high depth of read. Substitution of the estimated child genotype into equation 5 results in the complete equation (6) for the fetal fraction estimate.

$$\hat{f} = \arg \max_f \left[\prod_{s \in S_0} \left(\max_{r_s \in R_0(f)} P(n_{as}, n_{bs}|r_s) \right) \prod_{s \in S_1} \left(\max_{r_s \in R_1(f)} P(n_{as}, n_{bs}|r_s) \right) \right] \quad (6)$$

[0751] The fetal fraction must be in the range $[0, 1]$ and so the optimization can be easily implemented by a constrained one-dimensional search.

[0752] In the presence of low depth of read or high noise level, it may be preferable not to assume the maximum likelihood genotype, which may result in artificially high confidences. Another method would be to sum over the possible genotypes at each SNP, resulting in the following expression (7) for $P(n_a, n_b|f)$ for a SNP in S_0 . The prior probability $P(r)$ could be

assumed uniform over $R_0(f)$, or could be based on population frequencies. The extension to group S_1 is trivial.

$$P(n_a, n_b|f) = \sum_{r \in R_0(f)} P(n_a, n_b|r)P(r) \quad (7)$$

[0753] In some embodiments the probabilities may be derived as follows. A confidence can be calculated from the data likelihoods of the two hypotheses H_t and H_r . The likelihood of each hypothesis is derived based on the response model, the estimated fetal fraction, the mother genotypes, allele population frequencies, and the plasma allele counts.

[0754] Define the following notation:

[0755] G_m, G_c true maternal and child genotypes

[0756] G_{af}, G_{tf} true genotypes of alleged father and of true father

[0757] $G(g_c, g_m, g_{tf}) = P(G_c = g_c | G_m = g_m, G_{tf} = g_{tf})$
inheritance probabilities

[0758] $P(g) = P(G_{tf} = g)$ population frequency of
genotype g at particular SNP

[0759] Assuming that the observation at each SNP is independent conditioned on the plasma allele ratio, the likelihood of a paternity hypothesis is the product of the likelihoods on the SNPs. The following equations derive the likelihood for a single SNP. Equation 8 is a general expression for the likelihood of any hypothesis h , which will then be broken down into the specific cases of H_t and H_r .

$$\begin{aligned} [0760] \quad P(n_a, n_b|h, G_m, G_{tf}, f) &= \sum_{g_c \in (0,0.5,1)} P(n_a, n_b | G_c = \\ g_c, G_m, G_{tf}, h, f) P(G_c = g_c | G_m, G_{tf}, h) \\ &= \sum_{g_c \in (0,0.5,1)} P(n_a, n_b | G_c = g_c, G_m, f) P(G_c = g_c | G_m, G_{tf}, h) \\ &= \sum_{g_c \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf}, h) \quad (8) \end{aligned}$$

[0761] In the case of H_t , the alleged father is the true father and the fetal genotypes are inherited from the maternal genotypes and alleged father genotypes according to equation 9.

$$\begin{aligned}
 [0762] \quad P(n_a, n_b | H_t, G_m, G_{tf}, f) &= \sum_{g_c \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf}, H_t) \\
 (9) \\
 &= \sum_{g_c \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) G(g_c, G_m, G_{tf})
 \end{aligned}$$

[0763] In the case of H_f , the alleged father is not the true father. The best estimate of the true father genotypes are given by the population frequencies at each SNP. Thus, the probabilities of child genotypes are determined by the known mother genotypes and the population frequencies, as in equation 10.

$$\begin{aligned}
 [0764] \quad P(n_a, n_b | H_t, G_m, G_{tf}, f) &= \sum_{g_c \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf}, H_f) \\
 &= \sum_{g_c \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m) \\
 [0765] &= \sum_{g_c \in (0,0.5,1)} \sum_{g_{tf} \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) P(G_c = g_c | G_m, G_{tf} = g_{tf}) P(G_{tf} = g_{tf}) \\
 [0766] &= \sum_{g_c \in (0,0.5,1)} \sum_{g_{tf} \in (0,0.5,1)} F(n_a, n_b, g_c, g_m, f) G(g_c, G_m, g_{tf}) P(g_{tf})
 \end{aligned}$$

[0767] The confidence C_p on correct paternity is calculated from the product over SNPs of the two likelihoods using Bayes rule (11).

$$[0767] \quad C_p = \frac{\prod_s P(n_{as}, n_{bs} | H_t, G_{ms}, G_{tf}, f)}{\prod_s P(n_{as}, n_{bs} | H_t, G_{ms}, G_{tf}, f) + \prod_s P(n_{as}, n_{bs} | H_f, G_{ms}, G_{tf}, f)} \quad (11)$$

[0768] *Exemplary Methods for Identifying and Analyzing Multiple Pregnancies*

[0769] In some embodiments, any of the methods of the present invention are used to detect the presence of a multiple pregnancy, such as a twin pregnancy, where at least one of the fetuses is genetically different from at least one other fetus. In some embodiments, fraternal twins are

identified based on the presence of two fetus with different allele, different allele ratios, or different allele distributions at some (or all) of the tested loci. In some embodiments, fraternal twins are identified by determining the expected allele ratio at each locus (such as SNP loci) for two fetuses that may have the same or different fetal fractions in the sample (such as a plasma sample). In some embodiments, the likelihood of a particular pair of fetal fractions (where f_1 is the fetal fraction for fetus 1, and f_2 is the fetal fraction for fetus 2) is calculated by considering some or all of the possible genotypes of the two fetuses, conditioned on the mother's genotype and genotype population frequencies. The mixture of two fetal and one maternal genotype, combined with the fetal fractions, determine the expected allele ratio at a SNP. For example, if the mother is AA, fetus 1 is AA, and fetus 2 is AB, the overall fraction of B allele at the SNP is one-half of f_2 . The likelihood calculation asks how well all of the SNPs together match the expected allele ratios based on all of the possible combinations of fetal genotypes. The fetal fraction pair (f_1 , f_2) that best matches the data is selected. It is not necessary to calculate specific genotypes of the fetuses; instead, one can, for example, consider all of the possible genotypes in a statistical combination. In some embodiments, if the method does not distinguish between singleton and identical twins, an ultrasound can be performed to determine whether there is a singleton or identical twin pregnancy. If the ultrasound detects a twin pregnancy it can be assumed that the pregnancy is an identical twin pregnancy because a fraternal twin pregnancy would have been detected based on the SNP analysis discussed above.

[0770] In some embodiments, a pregnant mother is known to have a multiple pregnancy (such as a twin pregnancy) based on prior testing, such as an ultrasound. Any of the methods of the present invention can be used to determine whether the multiple pregnancy includes identical or fraternal twins. For example, the measured allele ratios can be compared to what would be expected for identical twins (the same allele ratios as a singleton pregnancy) or for fraternal twins (such as the calculation of allele ratios as described above). Some identical twins are monochorionic twins, which have a risk of twin-to-twin transfusion syndrome. Thus, twins determined to be identical twins using a method of the invention are desirably tested (such as by ultrasound) to determine if they are monochorionic twins, and if so, these twins can be monitored (such as bi-weekly ultrasounds from 16 weeks) for signs of twin-to-twin transfusion syndrome.

[0771] In some embodiments, any of the methods of the present invention are used to determine whether any of the fetuses in a multiple pregnancy, such as a twin pregnancy, are aneuploid. Aneuploidy testing for twins begins with the fetal fraction estimate. In some embodiments, the

fetal fraction pair (f_1 , f_2) that best matches the data is selected as described above. In some embodiments, a maximum likelihood estimate is performed for the parameter pair (f_1 , f_2) over the range of possible fetal fractions. In some embodiments, the range of f_2 is from 0 to f_1 because f_2 is defined as the smaller fetal fraction. Given a pair (f_1 , f_2), data likelihood is calculated from the allele ratios observed at a set of loci such as SNP loci. In some embodiments, the data likelihood reflects the genotypes of the mother, the father if available, population frequencies, and the resulting probabilities of fetal genotypes. In some embodiments, SNPs are assumed independent. The estimated fetal fraction pair is the one that produces the highest data likelihood. If f_2 is 0 then the data is best explained by only one set of fetal genotypes, indicating identical twins, where f_1 is the combined fetal fraction. Otherwise f_1 and f_2 are the estimates of the individual twin fetal fractions. Having established the best estimate of (f_1 , f_2), one can predict the overall fraction of B allele in the plasma for any combination of maternal and fetal genotypes, if desired. It is not necessary to assign individual sequence reads to the individual fetuses. Ploidy testing is performed using another maximum likelihood estimate which compares the data likelihood of two hypotheses. In some embodiments for identical twins, one considers the hypotheses (i) both twins are euploid, and (ii) both twins are trisomic. In some embodiments for fraternal twins, one considers the hypotheses (i) both twins are euploid and (ii) at least one twin is trisomic. The trisomy hypotheses for fraternal twins are based on the lower fetal fraction, since a trisomy in the twin with a higher fetal fraction would also be detected. Ploidy likelihoods are calculated using a method which predicts the expected number of reads at each targeted genome locus conditioned on either the disomy or trisomy hypothesis. There is no requirement for a disomy reference chromosome. The variance model for the expected number of reads takes into account the performance of individual target loci as well as the correlation between loci (see, for example, U.S. Serial No. 62/008,235, filed June 5, 2014, and U.S. Serial No. 62/032,785, filed August 4, 2014, which are each hereby incorporated by reference in its entirety). If the smaller twin has fetal fraction f_1 , our ability to detect a trisomy in that twin is equivalent to our ability to detect a trisomy in a singleton pregnancy at the same fetal fraction. This is because the part of the method that detects the trisomy in some embodiments does not depend on genotypes and does not distinguish between multiple or singleton pregnancy. It simply looks for an increased number of reads in accordance with the determined fetal fraction.

[0772] In some embodiments, the method includes detecting the presence of twins based on SNP loci (such as described above). If twins are detected, SPNs are used to determine the fetal fraction of each fetus (f1, f2) such as described above. In some embodiments, samples that have high confidence disomy calls are used to determine the amplification bias on a per-SNP basis. In some embodiments, these samples with high confidence disomy calls are analyzed in the same run as one or more samples of interest. In some embodiments, the amplification bias on a per-SNP basis is used to model the distribution of reads for one or more chromosomes or chromosome segments of interest such as chromosome 21 that are expected or the disomy hypothesis and the trisomy hypothesis given the lower of the two twin fetal fraction. The likelihood or probability of disomy or trisomy is calculated given the two models and the measured quantity of the chromosome or chromosome segment of interest.

[0773] In some embodiments, the threshold for a positive aneuploidy call (such as a trisomy call) is set based on the twin with the lower fetal fraction. This way, if the other twin is positive, or if both are positive, the total chromosome representation is definitely above the threshold.

[0774] *Maximum Likelihood Model using Percent Fetal Fraction*

[0775] Determining the ploidy status of a fetus by measuring the free floating DNA contained in maternal serum, or by measuring the genotypic material in any mixed sample, is a non-trivial exercise. There are a number of methods, for example, performing a read count analysis where the presumption is that if the fetus is trisomic at a particular chromosome, then the overall amount of DNA from that chromosome found in the maternal blood will be elevated with respect to a reference chromosome. One way to detect trisomy in such fetuses is to normalize the amount of DNA expected for each chromosome, for example, according to the number of SNPs in the analysis set that correspond to a given chromosome, or according to the number of uniquely mappable portions of the chromosome. Once the measurements have been normalized, any chromosomes for which the amount of DNA measured exceeds a certain threshold are determined to be trisomic. This approach is described in Fan, et al. PNAS, 2008; 105(42); pp. 16266–16271, and also in Chiu et al. BMJ 2011;342:c7401. In the Chiu et al. paper, the normalization was accomplished by calculating a Z score as follows:

[0776] Z
score for percentage chromosome 21 in test case = ((percentage chromosome 21 in test

case) – (mean percentage chromosome 21 in reference controls)) / (standard deviation of percentage chromosome 21 in reference controls).

[0777] These methods determine the ploidy status of the fetus using a single hypothesis rejection method. However, they suffer from some significant shortcomings. Since these methods for determining ploidy in the fetus are invariant according to the percentage of fetal DNA in the sample, they use one cut off value; the result of this is that the accuracies of the determinations are not optimal, and those cases where the percentage of fetal DNA in the mixture are relatively low will suffer the worst accuracies.

[0778] In an embodiment, a method of the present disclosure is used to determine the ploidy state of the fetus involves taking into account the fraction of fetal DNA in the sample. In another embodiment of the present disclosure, the method involves the use of maximum likelihood estimations. In an embodiment, a method of the present disclosure involves calculating the percent of DNA in a sample that is fetal or placental in origin. In an embodiment, the threshold for calling aneuploidy is adaptively adjusted based on the calculated percent fetal DNA. In some embodiments, the method for estimating the percentage of DNA that is of fetal origin in a mixture of DNA, comprises obtaining a mixed sample that comprises genetic material from the mother, and genetic material from the fetus, obtaining a genetic sample from the father of the fetus, measuring the DNA in the mixed sample, measuring the DNA in the father sample, and calculating the percentage of DNA that is of fetal origin in the mixed sample using the DNA measurements of the mixed sample, and of the father sample.

[0779] In an embodiment of the present disclosure, the fraction of fetal DNA, or the percentage of fetal DNA in the mixture can be measured. In some embodiments the fraction can be calculated using only the genotyping measurements made on the maternal plasma sample itself, which is a mixture of fetal and maternal DNA. In some embodiments the fraction may be calculated also using the measured or otherwise known genotype of the mother and/or the measured or otherwise known genotype of the father. In some embodiments the percent fetal DNA may be calculated using the measurements made on the mixture of maternal and fetal DNA along with the knowledge of the parental contexts. In an embodiment, the fraction of fetal DNA may be calculated using population frequencies to adjust the model on the probability on particular allele measurements.

[0780] In an embodiment of the present disclosure, a confidence may be calculated on the accuracy of the determination of the ploidy state of the fetus. In an embodiment, the confidence of the hypothesis of greatest likelihood (H_{major}) may be calculated as $(1 - H_{\text{major}}) / \Sigma(\text{all } H)$. It is possible to determine the confidence of a hypothesis if the distributions of all of the hypotheses are known. It is possible to determine the distribution of all of the hypotheses if the parental genotype information is known. It is possible to calculate a confidence of the ploidy determination if the knowledge of the expected distribution of data for the euploid fetus and the expected distribution of data for the aneuploid fetus are known. It is possible to calculate these expected distributions if the parental genotype data are known. In an embodiment one may use the knowledge of the distribution of a test statistic around a normal hypothesis and around an abnormal hypothesis to determine both the reliability of the call as well as refine the threshold to make a more reliable call. This is particularly useful when the amount and/or percent of fetal DNA in the mixture is low. It will help to avoid the situation where a fetus that is actually aneuploid is found to be euploid because a test statistic, such as the Z statistic does not exceed a threshold that is made based on a threshold that is optimized for the case where there is a higher percent fetal DNA.

[0781] In an embodiment, a method disclosed herein can be used to determine a fetal aneuploidy by determining the number of copies of maternal and fetal target chromosomes in a mixture of maternal and fetal genetic material. This method may entail obtaining maternal tissue comprising both maternal and fetal genetic material; in some embodiments this maternal tissue may be maternal plasma or a tissue isolated from maternal blood. This method may also entail obtaining a mixture of maternal and fetal genetic material from said maternal tissue by processing the aforementioned maternal tissue. This method may entail distributing the genetic material obtained into a plurality of reaction samples, to randomly provide individual reaction samples that comprise a target sequence from a target chromosome and individual reaction samples that do not comprise a target sequence from a target chromosome, for example, performing high throughput sequencing on the sample. This method may entail analyzing the target sequences of genetic material present or absent in said individual reaction samples to provide a first number of binary results representing presence or absence of a presumably euploid fetal chromosome in the reaction samples and a second number of binary results representing presence or absence of a possibly aneuploid fetal chromosome in the reaction samples. Either of the number of binary results may be calculated, for example, by way of an informatics technique that counts sequence

reads that map to a particular chromosome, to a particular region of a chromosome, to a particular locus or set of loci. This method may involve normalizing the number of binary events based on the chromosome length, the length of the region of the chromosome, or the number of loci in the set. This method may entail calculating an expected distribution of the number of binary results for a presumably euploid fetal chromosome in the reaction samples using the first number. This method may entail calculating an expected distribution of the number of binary results for a presumably aneuploid fetal chromosome in the reaction samples using the first number and an estimated fraction of fetal DNA found in the mixture, for example, by multiplying the expected read count distribution of the number of binary results for a presumably euploid fetal chromosome by $(1 + n/2)$ where n is the estimated fetal fraction. In some embodiments, the sequence reads may be treated at probabilistic mappings rather than binary results; this method would yield higher accuracies, but require more computing power. The fetal fraction may be estimated by a plurality of methods, some of which are described elsewhere in this disclosure. This method may involve using a maximum likelihood approach to determine whether the second number corresponds to the possibly aneuploid fetal chromosome being euploid or being aneuploid. This method may involve calling the ploidy status of the fetus to be the ploidy state that corresponds to the hypothesis with the maximum likelihood of being correct given the measured data.

[0782] Note that the use of a maximum likelihood model may be used to increase the accuracy of any method that determines the ploidy state of a fetus. Similarly, a confidence may be calculated for any method that determines the ploidy state of the fetus. The use of a maximum likelihood model would result in an improvement of the accuracy of any method where the ploidy determination is made using a single hypothesis rejection technique. A maximum likelihood model may be used for any method where a likelihood distribution can be calculated for both the normal and abnormal cases. The use of a maximum likelihood model implies the ability to calculate a confidence for a ploidy call.

[0783] *Further Discussion of the Method*

[0784] In an embodiment, a method disclosed herein utilizes a quantitative measure of the number of independent observations of each allele at a polymorphic locus, where this does not involve calculating the ratio of the alleles. This is different from methods, such as some

microarray based methods, which provide information about the ratio of two alleles at a locus but do not quantify the number of independent observations of either allele. Some methods known in the art can provide quantitative information regarding the number of independent observations, but the calculations leading to the ploidy determination utilize only the allele ratios, and do not utilize the quantitative information. To illustrate the importance of retaining information about the number of independent observations consider the sample locus with two alleles, A and B. In a first experiment twenty A alleles and twenty B alleles are observed, in a second experiment 200 A alleles and 200 B alleles are observed. In both experiments the ratio ($A/(A+B)$) is equal to 0.5, however the second experiment conveys more information than the first about the certainty of the frequency of the A or B allele. The instant method, rather than utilizing the allele ratios, uses the quantitative data to more accurately model the most likely allele frequencies at each polymorphic locus.

[0785] In an embodiment, the instant methods build a genetic model for aggregating the measurements from multiple polymorphic loci to better distinguish trisomy from disomy and also to determine the type of trisomy. Additionally, the instant method incorporates genetic linkage information to enhance the accuracy of the method. This is in contrast to some methods known in the art where allele ratios are averaged across all polymorphic loci on a chromosome. The method disclosed herein explicitly models the allele frequency distributions expected in disomy as well as and trisomy resulting from nondisjunction during meiosis I, nondisjunction during meiosis II, and nondisjunction during mitosis early in fetal development. To illustrate why this is important, if there were no crossovers nondisjunction during meiosis I would result a trisomy in which two different homologs were inherited from one parent; nondisjunction during meiosis II or during mitosis early in fetal development would result in two copies of the same homolog from one parent. Each scenario results in different expected allele frequencies at each polymorphic locus and also at all physically linked loci (*i.e.* loci on the same chromosome) considered jointly. Crossovers, which result in the exchange of genetic material between homologs, make the inheritance pattern more complex, but the instant method accommodates for this by using genetic linkage information, *i.e.* recombination rate information and the physical distance between loci. To better distinguish between meiosis I nondisjunction and meiosis II or mitotic nondisjunction the instant method incorporates into the model an increasing probability of crossover as the distance from the centromere increases. Meiosis II and mitotic nondisjunction can distinguished by the fact that mitotic nondisjunction typically results in

identical or nearly identical copies of one homolog while the two homologs present following a meiosis II nondisjunction event often differ due to one or more crossovers during gametogenesis. [0786] In an embodiment, a method of the present disclosure may not determine the haplotypes of the parents if disomy is assumed. In an embodiment, in case of trisomy, the instant method can make a determination about the haplotypes of one or both parents by using the fact that plasma takes two copies from one parent, and parent phase information can be determined by noting which two copies have been inherited from the parent in question. In particular, a child can inherit either two of the same copies of the parent (matched trisomy) or both copies of the parent (unmatched trisomy). At each SNP one can calculate the likelihood of the matched trisomy and of the unmatched trisomy. A ploidy calling method that does not use the linkage model accounting for crossovers would calculate the overall likelihood of the trisomy as a simple weighted average of the matched and unmatched trisomies over all chromosomes. However, due to the biological mechanisms that result in disjunction error and crossing over, trisomy can change from matched to unmatched (and vice versa) on a chromosome only if a crossover occurs. The instant method probabilistically takes into account the likelihood of crossover, resulting in ploidy calls that are of greater accuracy than those methods that do not.

[0787] In an embodiment, a reference chromosome is used to determine the child fraction and noise level amount or probability distribution. In an embodiment, the child fraction, noise level, and/or probability distribution is determined using only the genetic information available from the chromosome whose ploidy state is being determined. The instant method works without the reference chromosome, as well as without fixing the particular child fraction or noise level. This is a significant improvement and point of differentiation from methods known in the art where genetic data from a reference chromosome is necessary to calibrate the child fraction and chromosome behavior.

[0788] In an embodiment where a reference chromosome is not needed to determine the fetal fraction, determining the hypothesis is done as follows:

$$H^* = \underset{H}{\operatorname{argmax}} \operatorname{LIK}(D|H) * \operatorname{priorprob}(H)$$

[0789] With the algorithm with reference chromosome, one typically assumes that the reference chromosome is a disomy, and then one may either (a) fix the most likely child fraction and random noise level N based on this assumption and reference chromosome data :

$$[cfr^*, N^*] = \underset{cfr, N}{\operatorname{argmax}} \operatorname{LIK}(D(\text{ref. chrom})|H11, cfr, N)$$

[0790] And then reduce

$$\operatorname{LIK}(D|H) = \operatorname{LIK}(D|H, cfr^*, N^*)$$

[0790] or (b) estimate the child fraction and noise level distribution based on this assumption and reference chromosome data. In particular, one would not fix just one value for cfr and N, but assign probability $p(cfr, N)$ for the wider range of possible cfr, N values:

$$p(cfr, N) \sim \operatorname{LIK}(D(\text{ref. chrom})|H11, cfr, N) * \text{priorprob}(cfr, N)$$

[0791] where $\text{priorprob}(cfr, N)$ is the prior probability of particular child fraction and noise level, determined by prior knowledge and experiments. If desired, just uniform over the range of cfr, N. One may then write:

$$\operatorname{LIK}(D|H) = \sum_{cfr, N} \operatorname{LIK}(D|H, cfr, N) * p(cfr, N)$$

[0792] Both methods above give good results.

[0793] Note that in some instances using a reference chromosome is not desirable, possible or feasible. In such a case, it is possible to derive the best ploidy call for each chromosome separately. In particular:

$$\operatorname{LIK}(D|H) = \sum_{cfr, N} \operatorname{LIK}(D|H, cfr, N) * p(cfr, N|H)$$

[0794] $p(cfr, N|H)$ may be determined as above, for each chromosome separately, assuming hypothesis H, not just for the reference chromosome assuming disomy. It is possible, using this method, to keep both noise and child fraction parameters fixed, fix either of the parameters, or keep both parameters in probabilistic form for each chromosome and each hypothesis.

[0795] Measurements of DNA are noisy and/or error prone, especially measurements where the amount of DNA is small, or where the DNA is mixed with contaminating DNA. This noise results in less accurate genotypic data, and less accurate ploidy calls. In some embodiments, platform modeling or some other method of noise modeling may be used to counter the deleterious effects of noise on the ploidy determination. The instant method uses a joint model of both channels, which accounts for the random noise due to the amount of input DNA, DNA quality, and/or protocol quality.

[0796] This is in contrast to some methods known in the art where the ploidy determinations are made using the ratio of allele intensities at a locus. This method precludes accurate SNP noise modeling. In particular, errors in the measurements typically do not specifically depend on the measured channel intensity ratio, which reduces the model to using one-dimensional information. Accurate modeling of noise, channel quality and channel interaction requires a two-dimensional joint model, which can *not* be modeled using allele ratios.

[0797] In particular, projecting two channel information to the ratio r where $f(x,y)$ is $r = x/y$, does not lend itself to accurate channel noise and bias modeling. Noise on a particular SNP is not a function of the ratio, *i.e.* $\text{noise}(x,y) \neq f(x,y)$ but is in fact a joint function of both channels. For example, in the binomial model, noise of the measured ratio has a variance of $r(1-r)/(x+y)$ which is not a function purely of r . In such a model, where any channel bias or noise is included, suppose that on SNP i , the observed channel X value is $x=a_iX+b_i$, where X is the true channel value, b_i is the extra channel bias and random noise. Similarly, suppose that $y=c_iY+d_i$. The observed ratio $r=x/y$ can not accurately predict the true ratio X/Y or model the leftover noise, since $(a_iX+b_i)/(c_iY+d_i)$ is not a function of X/Y .

[0798] The method disclosed herein describes an effective way to model noise and bias using joint binomial distributions of all of the measurement channels individually. Relevant equations may be found elsewhere in the document in sections which speaks of per SNP consistent bias, $P(\text{good})$ and $P(\text{ref}|\text{bad})$, $P(\text{mut}|\text{bad})$ which effectively adjust SNP behavior. In an embodiment, a method of the present disclosure uses a BetaBinomial distribution, which avoids the limiting practice of relying on the allele ratios only, but instead models the behavior based on both channel counts.

[0799] In an embodiment, a method disclosed herein can call the ploidy of a gestating fetus from genetic data found in maternal plasma by using all available measurements. In an embodiment, a method disclosed herein can call the ploidy of a gestating fetus from genetic data found in maternal plasma by using the measurements from only a subset of parental contexts. Some methods known in the art only use measured genetic data where the parental context is from the AA|BB context, that is, where the parents are both homozygous at a given locus, but for a different allele. One problem with this method is that a small proportion of polymorphic loci are from the AA|BB context, typically less than 10%. In an embodiment of a method disclosed herein, the method does not use genetic measurements of the maternal plasma made at loci where

the parental context is AA|BB. In an embodiment, the instant method uses plasma measurements for only those polymorphic loci with the AA|AB, AB|AA, and AB|AB parental context.

[0800] Some methods known in the art involve averaging allele ratios from SNPs in the AA|BB context, where both parent genotypes are present, and claim to determine the ploidy calls from the average allele ratio on these SNPs. This method suffers from significant inaccuracy due differential SNP behavior. Note that this method assumes that have both parent genotypes are known. In contrast, in some embodiments, the instant method uses a joint channel distribution model that does not assume the presence of either of the parents, and does not assume the uniform SNP behavior. In some embodiments, the instant method accounts for the different SNP behavior/weighing. In some embodiments, the instant method does not require the knowledge of one or both parental genotypes. An example of how the instant method may accomplish this follows:

[0801] In some embodiments, the log likelihood of a hypothesis may be determined on a per SNP basis. On a particular SNP i , assuming fetal ploidy hypothesis H and percent fetal DNA cf , the log likelihood of observed data D is defined as:

$$LIK(D|H, i) = \log P(D|H, cf, i) = \log \left(\sum_{m,f,c} P(D|m, f, c, H, cf, i) P(c|m, f, H) P(m|i) P(f|i) \right)$$

[0802] where m are possible true mother genotypes, f are possible true father genotypes, where $m, f \in \{AA, AB, BB\}$, and where c are possible child genotypes given the hypothesis H . In particular, for monosomy $c \in \{A, B\}$, for disomy $c \in \{AA, AB, BB\}$, for trisomy $c \in \{AAA, AAB, ABB, BBB\}$. Note that including parental genotypic data typically results in more accurate ploidy determinations, however, parental genotypic data is not necessary for the instant method to work well.

[0803] Some methods known in the art involve averaging allele ratios from SNPs where the mother is homozygous but a different allele is measured in the plasma (either AA|AB or AA|BB contexts), and claim to determine the ploidy calls from the average allele ratio on these SNPs. This method is intended for cases where the paternal genotype is not available. Note that it is questionable how accurately one can claim that plasma is heterozygous on a particular SNP without the presence of homozygous and opposite father BB: for cases with low child fraction, what looks like presence of B allele could be just presence of noise; additionally, what looks like no B present could be simple allele drop out of the fetal measurements. Even in a case where

one can actually determine heterozygosity of the plasma, this method will not be able to distinguish paternal trisomies. In particular, for SNPs where mother is AA, and where some B is measured in the plasma, if the father is GG, the resulting child genotype is AGG, resulting in an average ratio of 33% A (for child fraction=100%). But in the case where the father is AG, the resulting child genotype could be AGG for matched trisomy, contributing to the 33% A ratio, or AAG for unmatched trisomy, drawing the average ratio more toward 66% A. Given that many trisomies are on chromosomes with crossovers, the overall chromosome can have anywhere between no unmatched trisomy and all unmatched trisomy, this ratio can vary anywhere between 33-66%. For a plain disomy, the ratio should be around 50%. Without the use of a linkage model or an accurate error model of the average, this method would miss many cases of paternal trisomy. In contrast, the method disclosed herein assigns parental genotype probabilities for each parental genotypic candidate, based on available genotypic information and population frequency, and does not explicitly require parental genotypes. Additionally, the method disclosed herein is able to detect trisomy even in the absence or presence of parent genotypic data, and can compensate by identifying the points of possible crossovers from matched to unmatched trisomy using a linkage model.

[0804] Some methods known in the art claim a method for averaging allele ratios from SNPs where neither the maternal or paternal genotype is known, and for determining the ploidy calls from average ratio on these SNPs. However, a method to accomplish these ends is not disclosed. The method disclosed herein is able to make accurate ploidy calls in such a situation, and the reduction to practice is disclosed elsewhere in this document, using a joint probability maximum likelihood method and optionally utilizes SNP noise and bias models, as well as a linkage model.

[0805] Some methods known in the art involve averaging allele ratios and claim to determine the ploidy calls from the average allele ratio at one or a few SNPs. However, such methods do not utilize the concept of linkage. The methods disclosed herein do not suffer from these drawbacks.

[0806] *Using Sequence Length as a Prior to Determine the Origin of DNA*

[0807] It has been reported that the distribution of length of sequences differ for maternal and fetal DNA, with fetal generally being shorter. In an embodiment of the present disclosure, it is possible to use previous knowledge in the form of empirical data, and construct prior distribution

for expected length of both mother ($P(X| \text{maternal})$) and fetal DNA ($P(X| \text{fetal})$). Given new unidentified DNA sequence of length x , it is possible to assign a probability that a given sequence of DNA is either maternal or fetal DNA, based on prior likelihood of x given either maternal or fetal. In particular if $P(x|\text{maternal}) > P(x|\text{fetal})$, then the DNA sequence can be classified as maternal, with $P(x|\text{maternal}) = P(x|\text{maternal}) / [P(x|\text{maternal}) + P(x|\text{fetal})]$, and if $p(x|\text{maternal}) < p(x|\text{fetal})$, then the DNA sequence can be classified as fetal, $P(x|\text{fetal}) = P(x|\text{fetal}) / [P(x|\text{maternal}) + P(x|\text{fetal})]$. In an embodiment of the present disclosure, a distributions of maternal and fetal sequence lengths can be determined that is specific for that sample by considering the sequences that can be assigned as maternal or fetal with high probability, and then that sample specific distribution can be used as the expected size distribution for that sample.

[0808] *Variable Read Depth to Minimize Sequencing Cost*

[0809] In many clinical trials concerning a diagnostic, for example, in Chiu et al. BMJ 2011;342:c7401, a protocol with a number of parameters is set, and then the same protocol is executed with the same parameters for each of the patients in the trial. In the case of determining the ploidy status of a fetus gestating in a mother using sequencing as a method to measure genetic material one pertinent parameter is the number of reads. The number of reads may refer to the number of actual reads, the number of intended reads, fractional lanes, full lanes, or full flow cells on a sequencer. In these studies, the number of reads is typically set at a level that will ensure that all or nearly all of the samples achieve the desired level of accuracy. Sequencing is currently an expensive technology, a cost of roughly \$200 per 5 mappable million reads, and while the price is dropping, any method which allows a sequencing based diagnostic to operate at a similar level of accuracy but with fewer reads will necessarily save a considerable amount of money.

[0810] The accuracy of a ploidy determination is typically dependent on a number of factors, including the number of reads and the fraction of fetal DNA in the mixture. The accuracy is typically higher when the fraction of fetal DNA in the mixture is higher. At the same time, the accuracy is typically higher if the number of reads is greater. It is possible to have a situation with two cases where the ploidy state is determined with comparable accuracies wherein the first case has a lower fraction of fetal DNA in the mixture than the second, and more reads were

sequenced in the first case than the second. It is possible to use the estimated fraction of fetal DNA in the mixture as a guide in determining the number of reads necessary to achieve a given level of accuracy.

[0811] In an embodiment of the present disclosure, a set of samples can be run where different samples in the set are sequenced to different reads depths, wherein the number of reads run on each of the samples is chosen to achieve a given level of accuracy given the calculated fraction of fetal DNA in each mixture. In an embodiment of the present disclosure, this may entail making a measurement of the mixed sample to determine the fraction of fetal DNA in the mixture; this estimation of the fetal fraction may be done with sequencing, it may be done with TAQMAN, it may be done with qPCR, it may be done with SNP arrays, it may be done with any method that can distinguish different alleles at a given loci. The need for a fetal fraction estimate may be eliminated by including hypotheses that cover all or a selected set of fetal fractions in the set of hypotheses that are considered when comparing to the actual measured data. After the fraction fetal DNA in the mixture has been determined, the number of sequences to be read for each sample may be determined.

[0812] In an embodiment of the present disclosure, 100 pregnant women visit their respective OB's, and their blood is drawn into blood tubes with an anti-lycant and/or something to inactivate DNAase. They each take home a kit for the father of their gestating fetus who gives a saliva sample. Both sets of genetic materials for all 100 couples are sent back to the laboratory, where the mother blood is spun down and the buffy coat is isolated, as well as the plasma. The plasma comprises a mixture of maternal DNA as well as placentally derived DNA. The maternal buffy coat and the paternal blood is genotyped using a SNP array, and the DNA in the maternal plasma samples are targeted with SURESELECT hybridization probes. The DNA that was pulled down with the probes is used to generate 100 tagged libraries, one for each of the maternal samples, where each sample is tagged with a different tag. A fraction from each library is withdrawn, each of those fractions are mixed together and added to two lanes of a ILLUMINA HISEQ DNA sequencer in a multiplexed fashion, wherein each lane resulted in approximately 50 million mappable reads, resulting in approximately 100 million mappable reads on the 100 multiplexed mixtures, or approximately 1 million reads per sample. The sequence reads were used to determine the fraction of fetal DNA in each mixture. 50 of the samples had more than 15% fetal DNA in the mixture, and the 1 million reads were sufficient to determine the ploidy status of the fetuses with a 99.9% confidence.

[0813] Of the remaining mixtures, 25 had between 10 and 15% fetal DNA; a fraction of each of the relevant libraries prepped from these mixtures were multiplexed and run down one lane of the HISEQ generating an additional 2 million reads for each sample. The two sets of sequence data for each of the mixture with between 10 and 15% fetal DNA were added together, and the resulting 3 million reads per sample which were sufficient to determine the ploidy state of those fetuses with 99.9% confidence.

[0814] Of the remaining mixtures, 13 had between 6 and 10% fetal DNA; a fraction of each of the relevant libraries prepped from these mixtures were multiplexed and run down one lane of the HISEQ generating an additional 4 million reads for each sample. The two sets of sequence data for each of the mixture with between 6 and 10% fetal DNA were added together, and the resulting 5 million total reads per mixture which were sufficient to determine the ploidy state of those fetuses with 99.9% confidence.

[0815] Of the remaining mixtures, 8 had between 4 and 6% fetal DNA; a fraction of each of the relevant libraries prepped from these mixtures were multiplexed and run down one lane of the HISEQ generating an additional 6 million reads for each sample. The two sets of sequence data for each of the mixture with between 4 and 6% fetal DNA were added together, and the resulting 7 million total reads per mixture which were sufficient to determine the ploidy state of those fetuses with 99.9% confidence.

[0816] Of the remaining four mixtures, all of them had between 2 and 4% fetal DNA; a fraction of each of the relevant libraries prepped from these mixtures were multiplexed and run down one lane of the HISEQ generating an additional 12 million reads for each sample. The two sets of sequence data for each of the mixture with between 2 and 4% fetal DNA were added together, and the resulting 13 million total reads per mixture which were sufficient to determine the ploidy state of those fetuses with 99.9% confidence.

[0817] This method required six lanes of sequencing on a HISEQ machine to achieve 99.9% accuracy over 100 samples. If the same number of runs had been required for every sample, to ensure that every ploidy determination was made with a 99.9% accuracy, it would have taken 25 lanes of sequencing, and if a no-call rate or error rate of 4% was tolerated, it could have been achieved with 14 lanes of sequencing.

[0818] *Using Raw Genotyping Data*

[0819] There are a number of methods that can accomplish NPD using fetal genetic information measured on fetal DNA found in maternal blood. Some of these methods involve making measurements of the fetal DNA using SNP arrays, some methods involve untargeted sequencing, and some methods involve targeted sequencing. The targeted sequencing may target SNPs, it may target STRs, it may target other polymorphic loci, it may target non-polymorphic loci, or some combination thereof. Some of these methods may involve using a commercial or proprietary allele caller that calls the identity of the alleles from the intensity data that comes from the sensors in the machine doing the measuring. For example, the ILLUMINA INFINIUM system or the AFFYMETRIX GENECHIP microarray system involves beads or microchips with attached DNA sequences that can hybridize to complementary segments of DNA; upon hybridization, there is a change in the fluorescent properties of the sensor molecule that can be detected. There are also sequencing methods, for example the ILLUMINA SOLEXA GENOME SEQUENCER or the ABI SOLID GENOME SEQUENCER, wherein the genetic sequence of fragments of DNA are sequenced; upon extension of the strand of DNA complementary to the strand being sequenced, the identity of the extended nucleotide is typically detected via a fluorescent or radio tag appended to the complementary nucleotide. In all of these methods the genotypic or sequencing data is typically determined on the basis of fluorescent or other signals, or the lack thereof. These systems are typically combined with low level software packages that make specific allele calls (secondary genetic data) from the analog output of the fluorescent or other detection device (primary genetic data). For example, in the case of a given allele on a SNP array, the software will make a call, for example, that a certain SNP is present or not present if the fluorescent intensity is measure above or below a certain threshold. Similarly, the output of a sequencer is a chromatogram that indicates the level of fluorescence detected for each of the dyes, and the software will make a call that a certain base pair is A or T or C or G. High throughput sequencers typically make a series of such measurements, called a read, that represents the most likely structure of the DNA sequence that was sequenced. The direct analog output of the chromatogram is defined here to be the primary genetic data, and the base pair / SNP calls made by the software are considered here to be the secondary genetic data. In an embodiment, primary data refers to the raw intensity data that is the unprocessed output of a genotyping platform, where the genotyping platform may refer to a SNP array, or to a sequencing platform. The secondary genetic data refers to the processed genetic data, where an allele call

has been made, or the sequence data has been assigned base pairs, and/or the sequence reads have been mapped to the genome.

[0820] Many higher level applications take advantage of these allele calls, SNP calls and sequence reads, that is, the secondary genetic data, that the genotyping software produces. For example, DNA NEXUS, ELAND or MAQ will take the sequencing reads and map them to the genome. For example, in the context of non-invasive prenatal diagnosis, complex informatics, such as PARENTAL SUPPORT™, may leverage a large number of SNP calls to determine the genotype of an individual. Also, in the context of preimplantation genetic diagnosis, it is possible to take a set of sequence reads that are mapped to the genome, and by taking a normalized count of the reads that are mapped to each chromosome, or section of a chromosome, it may be possible to determine the ploidy state of an individual. In the context of non-invasive prenatal diagnosis it may be possible to take a set of sequence reads that have been measured on DNA present in maternal plasma, and map them to the genome. One may then take a normalized count of the reads that are mapped to each chromosome, or section of a chromosome, and use that data to determine the ploidy state of an individual. For example, it may be possible to conclude that those chromosomes that have a disproportionately large number of reads are trisomic in the fetus that is gestating in the mother from which the blood was drawn.

[0821] However, in reality, the initial output of the measuring instruments is an analog signal. When a certain base pair is called by the software that is associated with the sequencing software, for example the software may call the base pair a T, in reality the call is the call that the software believes to be most likely. In some cases, however, the call may be of low confidence, for example, the analog signal may indicate that the particular base pair is only 90% likely to be a T, and 10% likely to be an A. In another example, the genotype calling software that is associated with a SNP array reader may call a certain allele to be G. However, in reality, the underlying analog signal may indicate that it is only 70% likely that the allele is G, and 30% likely that the allele is T. In these cases, when the higher level applications use the genotype calls and sequence calls made by the lower level software, they are losing some information. That is, the primary genetic data, as measured directly by the genotyping platform, may be messier than the secondary genetic data that is determined by the attached software packages, but it contains more information. In mapping the secondary genetic data sequences to the genome, many reads are thrown out because some bases are not read with enough clarity and or mapping is not clear. When the primary genetic data sequence reads are used, all or many of

those reads that may have been thrown out when first converted to secondary genetic data sequence read can be used by treating the reads in a probabilistic manner.

[0822] In an embodiment of the present disclosure, the higher level software does not rely on the allele calls, SNP calls, or sequence reads that are determined by the lower level software. Instead, the higher level software bases its calculations on the analog signals directly measured from the genotyping platform. In an embodiment of the present disclosure, an informatics based method such as PARENTAL SUPPORT™ is modified so that its ability to reconstruct the genetic data of the embryo / fetus / child is engineered to directly use the primary genetic data as measured by the genotyping platform. In an embodiment of the present disclosure, an informatics based method such as PARENTAL SUPPORT™ is able to make allele calls, and/or chromosome copy number calls using primary genetic data, and not using the secondary genetic data. In an embodiment of the present disclosure, all genetic calls, SNPs calls, sequence reads, sequence mapping is treated in a probabilistic manner by using the raw intensity data as measured directly by the genotyping platform, rather than converting the primary genetic data to secondary genetic calls. In an embodiment, the DNA measurements from the prepared sample used in calculating allele count probabilities and determining the relative probability of each hypothesis comprise primary genetic data.

[0823] In some embodiments, the method can increase the accuracy of genetic data of a target individual which incorporates genetic data of at least one related individual, the method comprising obtaining primary genetic data specific to a target individual's genome and genetic data specific to the genome(s) of the related individual(s), creating a set of one or more hypotheses concerning possibly which segments of which chromosomes from the related individual(s) correspond to those segments in the target individual's genome, determining the probability of each of the hypotheses given the target individual's primary genetic data and the related individual(s)'s genetic data, and using the probabilities associated with each hypothesis to determine the most likely state of the actual genetic material of the target individual. In some embodiments, the method can determining the number of copies of a segment of a chromosome in the genome of a target individual, the method comprising creating a set of copy number hypotheses about how many copies of the chromosome segment are present in the genome of a target individual, incorporating primary genetic data from the target individual and genetic information from one or more related individuals into a data set, estimating the characteristics of the platform response associated with the data set, where the platform response may vary

from one experiment to another, computing the conditional probabilities of each copy number hypothesis, given the data set and the platform response characteristics, and determining the copy number of the chromosome segment based on the most probable copy number hypothesis. In an embodiment, a method of the present disclosure can determine a ploidy state of at least one chromosome in a target individual, the method comprising obtaining primary genetic data from the target individual and from one or more related individuals, creating a set of at least one ploidy state hypothesis for each of the chromosomes of the target individual, using one or more expert techniques to determine a statistical probability for each ploidy state hypothesis in the set, for each expert technique used, given the obtained genetic data, combining, for each ploidy state hypothesis, the statistical probabilities as determined by the one or more expert techniques, and determining the ploidy state for each of the chromosomes in the target individual based on the combined statistical probabilities of each of the ploidy state hypotheses. In an embodiment, a method of the present disclosure can determine an allelic state in a set of alleles, in a target individual, and from one or both parents of the target individual, and optionally from one or more related individuals, the method comprising obtaining primary genetic data from the target individual, and from the one or both parents, and from any related individuals, creating a set of at least one allelic hypothesis for the target individual, and for the one or both parents, and optionally for the one or more related individuals, where the hypotheses describe possible allelic states in the set of alleles, determining a statistical probability for each allelic hypothesis in the set of hypotheses given the obtained genetic data, and determining the allelic state for each of the alleles in the set of alleles for the target individual, and for the one or both parents, and optionally for the one or more related individuals, based on the statistical probabilities of each of the allelic hypotheses.

[0824] In some embodiments, the genetic data of the mixed sample may comprise sequence data wherein the sequence data may not uniquely map to the human genome. In some embodiments, the genetic data of the mixed sample may comprise sequence data wherein the sequence data maps to a plurality of locations in the genome, wherein each possible mapping is associated with a probability that the given mapping is correct. In some embodiments, the sequence reads are not assumed to be associated with a particular position in the genome. In some embodiments, the sequence reads are associated with a plurality of positions in the genome, and an associated probability belonging to that position.

[0825] *Counting Method to Determine Chromosome Copy Number*

[0826] In one aspect, the invention features methods of testing for an abnormal distribution of a fetal chromosome by comparing the number of sequence tags that align to different chromosomes (*see, e.g.*, U.S. Pat. No. 8,296,076, filed April 20, 2012, which is hereby incorporated by reference in its entirety). As is known in the art, the term "sequence tag" refers to a relatively short (*e.g.*, 15-100) nucleic acid sequence that can be used to identify a certain larger sequence, *e.g.*, be mapped to a chromosome or genomic region or gene. In some embodiments, the method involves (i) contacting a sample that includes a mixture of maternal and fetal DNA with a library of primers that simultaneously hybridize to at least 25; 50; 75; 100; 300; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 15,000; 19,000; 20,000; 25,000; 27,000; 28,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci to produce a reaction mixture; wherein the target loci are from a plurality of different chromosomes; and wherein the plurality of different chromosomes comprise at least one first chromosome suspected of having an abnormal distribution in the sample and at least one second chromosome presumed to be normally distributed in the sample; (ii) subjecting the reaction mixture to primer extension reaction conditions to produce amplified products; (iii) sequencing the amplified products to obtain a plurality of sequence tags aligning to the target loci; wherein the sequence tags are of sufficient length to be assigned to a specific target locus; (iv) assigning on a computer the plurality of sequence tags to their corresponding target loci; (v) determining on a computer a number of sequence tags aligning to the target loci of the first chromosome and a number of sequence tags aligning to the target loci of the second chromosome; and (vi) comparing the numbers from step (v) to determine the presence or absence of an abnormal distribution of the first chromosome.

[0827] In one aspect, the invention provides methods for detecting the presence or absence of a fetal aneuploidy by comparing the relative frequency of target amplicons between chromosomes (*see, e.g.*, PCT Publ. No. WO 2012/103031, filed Jan. 23, 2012, which is hereby incorporated by reference in its entirety). In some embodiments, the method involves (i) contacting a sample that includes a mixture of maternal and fetal DNA with a library of primers that simultaneously hybridize to at least 25; 50; 75; 100; 300; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 15,000; 19,000; 20,000; 25,000; 27,000; 28,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different non-polymorphic target loci to produce a reaction mixture;

wherein the target loci are from a plurality of different chromosomes; (ii) subjecting the reaction mixture to primer extension reaction conditions to produce amplified products that includes target amplicons; (iii) quantifying on a computer a relative frequency of the target amplicons from the first and second chromosomes of interest; (iv) comparing on a computer the relative frequency of the target amplicons from the first and second chromosomes of interest; and (v) identifying the presence or absence of an aneuploidy based on the compared relative frequencies of the first and second chromosome of interest. In some embodiments, the first chromosome is a chromosome suspected of being euploid. In some embodiments, the second chromosome is a chromosome suspected of being aneuploidy

[0828] *Combining Methods of Prenatal Diagnosis*

[0829] There are many methods that may be used for prenatal diagnosis or prenatal screening of aneuploidy or other genetic defects. Described elsewhere in this document, and in U.S. Utility Application Serial No. 11/603,406, filed November 28, 2006; U.S. Utility Application Serial No. 12/076,348, filed March 17, 2008, and PCT Application Serial No. PCT/S09/52730 is one such method that uses the genetic data of related individuals to increase the accuracy with which genetic data of a target individual, such as a fetus, is known, or estimated. Other methods used for prenatal diagnosis involve measuring the levels of certain hormones in maternal blood, where those hormones are correlated with various genetic abnormalities. An example of this is called the triple test, a test wherein the levels of several (commonly two, three, four or five) different hormones are measured in maternal blood. In a case where multiple methods are used to determine the likelihood of a given outcome, where none of the methods are definitive in and of themselves, it is possible to combine the information given by those methods to make a prediction that is more accurate than any of the individual methods. In the triple test, combining the information given by the three different hormones can result in a prediction of genetic abnormalities that is more accurate than the individual hormone levels may predict.

[0830] Disclosed herein is a method for making more accurate predictions about the genetic state of a fetus, specifically the possibility of genetic abnormalities in a fetus that comprises combining predictions of genetic abnormalities in a fetus where those predictions were made using a variety of methods. A “more accurate” method may refer to a method for diagnosing an abnormality that has a lower false negative rate at a given false positive rate. In a favored embodiment of the present disclosure, one or more of the predictions are made based on the

genetic data known about the fetus, where the genetic knowledge was determined using the PARENTAL SUPPORT™ method, that is, using genetic data of individual related to the fetus to determine the genetic data of the fetus with greater accuracy. In some embodiments the genetic data may include ploidy states of the fetus. In some embodiments, the genetic data may refer to a set of allele calls on the genome of the fetus. In some embodiments some of the predictions may have been made using the triple test. In some embodiments, some of the predictions may have been made using measurements of other hormone levels in maternal blood. In some embodiments, predictions made by methods considered diagnoses may be combined with predictions made by methods considered screening. In some embodiments, the method involves measuring maternal blood levels of alpha-fetoprotein (AFP). In some embodiments, the method involves measuring maternal blood levels of unconjugated estriol (UE₃). In some embodiments, the method involves measuring maternal blood levels of beta human chorionic gonadotropin (beta-hCG). In some embodiments, the method involves measuring maternal blood levels of invasive trophoblast antigen (ITA). In some embodiments, the method involves measuring maternal blood levels of inhibin. In some embodiments, the method involves measuring maternal blood levels of pregnancy-associated plasma protein A (PAPP-A). In some embodiments, the method involves measuring maternal blood levels of other hormones or maternal serum markers. In some embodiments, some of the predictions may have been made using other methods. In some embodiments, some of the predictions may have been made using a fully integrated test such as one that combines ultrasound and blood test at around 12 weeks of pregnancy and a second blood test at around 16 weeks. In some embodiments, the method involves measuring the fetal nuchal translucency (NT). In some embodiments, the method involves using the measured levels of the aforementioned hormones for making predictions. In some embodiments the method involves a combination of the aforementioned methods.

[0831] There are many ways to combine the predictions, for example, one could convert the hormone measurements into a multiple of the median (MoM) and then into likelihood ratios (LR). Similarly, other measurements could be transformed into LRs using the mixture model of NT distributions. The LRs for NT and the biochemical markers could be multiplied by the age and gestation-related risk to derive the risk for various conditions, such as trisomy 21. Detection rates (DRs) and false-positive rates (FPRs) could be calculated by taking the proportions with risks above a given risk threshold.

[0832] In an embodiment, a method to call the ploidy state involves combining the relative

probabilities of each of the ploidy hypotheses determined using the joint distribution model and the allele count probabilities with relative probabilities of each of the ploidy hypotheses that are calculated using statistical techniques taken from other methods that determine a risk score for a fetus being trisomic, including but not limited to: a read count analysis, comparing heterozygosity rates, a statistic that is only available when parental genetic information is used, the probability of normalized genotype signals for certain parent contexts, a statistic that is calculated using an estimated fetal fraction of the first sample or the prepared sample, and combinations thereof.

[0833] Another method could involve a situation with four measured hormone levels, where the probability distribution around those hormones is known: $p(x_1, x_2, x_3, x_4|e)$ for the euploid case and $p(x_1, x_2, x_3, x_4|a)$ for the aneuploid case. Then one could measure the probability distribution for the DNA measurements, $g(y|e)$ and $g(y|a)$ for the euploid and aneuploid cases respectively. Assuming they are independent given the assumption of euploid/aneuploid, one could combine as $p(x_1, x_2, x_3, x_4|a)g(y|a)$ and $p(x_1, x_2, x_3, x_4|e)g(y|e)$ and then multiply each by the prior $p(a)$ and $p(e)$ given the maternal age. One could then choose the one that is highest.

[0834] In an embodiment, it is possible to evoke central limit theorem to assume distribution on $g(y|a \text{ or } e)$ is Gaussian, and measure mean and standard deviation by looking at multiple samples. In another embodiment, one could assume they are not independent given the outcome and collect enough samples to estimate the joint distribution $p(x_1, x_2, x_3, x_4|a \text{ or } e)$.

[0835] In an embodiment, the ploidy state for the target individual is determined to be the ploidy state that is associated with the hypothesis whose probability is the greatest. In some cases, one hypothesis will have a normalized, combined probability greater than 90%. Each hypothesis is associated with one, or a set of, ploidy states, and the ploidy state associated with the hypothesis whose normalized, combined probability is greater than 90%, or some other threshold value, such as 50%, 80%, 95%, 98%, 99%, or 99.9%, may be chosen as the threshold required for a hypothesis to be called as the determined ploidy state.

[0836] *DNA from Children from Previous Pregnancies in Maternal Blood*

[0837] One difficulty to non-invasive prenatal diagnosis is differentiating fetal cells from the current pregnancy from fetal cells from previous pregnancies. Some believe that genetic matter from prior pregnancies will go away after some time, but conclusive evidence has not been

shown. In an embodiment of the present disclosure, it is possible to determine fetal DNA present in the maternal blood of paternal origin (that is, DNA that the fetus inherited from the father) using the PARENTAL SUPPORT™ (PS) method, and the knowledge of the paternal genome. This method may utilize phased parental genetic information. It is possible to phase the parental genotype from unphased genotypic information using grandparental genetic data (such as measured genetic data from a sperm from the grandfather), or genetic data from other born children, or a sample of a miscarriage. One could also phase unphased genetic information by way of a HapMap-based phasing, or a haplotyping of paternal cells. Successful haplotyping has been demonstrated by arresting cells at phase of mitosis when chromosomes are tight bundles and using microfluidics to put separate chromosomes in separate wells. In another embodiment it is possible to use the phased parental haplotypic data to detect the presence of more than one homolog from the father, implying that the genetic material from more than one child is present in the blood. By focusing on chromosomes that are expected to be euploid in a fetus, one could rule out the possibility that the fetus was afflicted with a trisomy. Also, it is possible to determine if the fetal DNA is not from the current father, in which case one could use other methods such as the triple test to predict genetic abnormalities.

[0838] There may be other sources of fetal genetic material available via methods other than a blood draw. In the case of the fetal genetic material available in maternal blood, there are two main categories: (1) whole fetal cells, for example, nucleated fetal red blood cells or erythroblasts, and (2) free floating fetal DNA. In the case of whole fetal cells, there is some evidence that fetal cells can persist in maternal blood for an extended period of time such that it is possible to isolate a cell from a pregnant woman that contains the DNA from a child or fetus from a prior pregnancy. There is also evidence that the free floating fetal DNA is cleared from the system in a matter of weeks. One challenge is how to determine the identity of the individual whose genetic material is contained in the cell, namely to ensure that the measured genetic material is not from a fetus from a prior pregnancy. In an embodiment of the present disclosure, the knowledge of the maternal genetic material can be used to ensure that the genetic material in question is not maternal genetic material. There are a number of methods to accomplish this end, including informatics based methods such as PARENTAL SUPPORT™, as described in this document or any of the patents referenced in this document.

[0839] In an embodiment of the present disclosure, the blood drawn from the pregnant mother may be separated into a fraction comprising free floating fetal DNA, and a fraction comprising

nucleated red blood cells. The free floating DNA may optionally be enriched, and the genotypic information of the DNA may be measured. From the measured genotypic information from the free floating DNA, the knowledge of the maternal genotype may be used to determine aspects of the fetal genotype. These aspects may refer to ploidy state, and/or a set of allele identities. Then, individual nucleated red blood cells may be genotyped using methods described elsewhere in this document, and other referent patents, especially those mentioned in the first section of this document. The knowledge of the maternal genome would allow one to determine whether or not any given single blood cell is genetically maternal. And the aspects of the fetal genotype that were determined as described above would allow one to determine if the single blood cell is genetically derived from the fetus that is currently gestating. In essence, this aspect of the present disclosure allows one to use the genetic knowledge of the mother, and possibly the genetic information from other related individuals, such as the father, along with the measured genetic information from the free floating DNA found in maternal blood to determine whether an isolated nucleated cell found in maternal blood is either (a) genetically maternal, (b) genetically from the fetus currently gestating, or (c) genetically from a fetus from a prior pregnancy.

[0840] *Prenatal Sex Chromosome Aneuploidy Determination*

[0841] In methods known in the art, people attempting to determine the sex of a gestating fetus from the blood of the mother have used the fact that fetal free floating DNA (fffDNA) is present in the plasma of the mother. If one is able to detect Y-specific loci in the maternal plasma, this implies that the gestating fetus is a male. However, the lack of detection of Y-specific loci in the plasma does not always guarantee that the gestating fetus is a female when using methods known in the art, as in some cases the amount of fffDNA is too low to ensure that the Y-specific loci would be detected in the case of a male fetus.

[0842] Presented here is a novel method that does not require the measurement of Y-specific nucleic acids, that is, DNA that is from loci that are exclusively paternally derived. The Parental Support method, disclosed previously, uses crossover frequency data, parental genotypic data, and informatics techniques, to determine the ploidy state of a gestating fetus. The sex of a fetus is simply the ploidy state of the fetus at the sex chromosomes. A child that is XX is female, and XY is male. The method described herein is also able to determine the ploidy state of the fetus.

Note that sexing is effectively synonymous with ploidy determination of the sex chromosomes; in the case of sexing, an assumption is often made that the child is euploid, therefore there are fewer possible hypotheses.

[0843] The method disclosed herein involves looking at loci that are common to both the X and Y chromosome to create a baseline in terms of expected amount of fetal DNA present for a fetus. Then, those regions that are specific only to the X chromosome can be interrogated to determine if the fetus is female or male. In the case of a male, we expect to see less fetal DNA from loci that are specific to the X chromosome than from loci that are specific to both the X and the Y. In contrast, in female fetuses, we expect the amount of DNA for each of these groups to be the same. The DNA in question can be measured by any technique that can quantitate the amount of DNA present on a sample, for example, qPCR, SNP arrays, genotyping arrays, or sequencing. For DNA that is exclusively from an individual we would expect to see the following:

	DNA specific to X	DNA specific to X and Y	DNA specific to Y
Male (XY)	A	2A	A
Female (XX)	2A	2A	0

[0844] In the case of DNA from a fetus that is mixed with DNA from the mother, and where the fraction of fetal DNA in the mixture is F, and where the fraction of maternal DNA in the mixture is M, such that $F+M = 100\%$, we would expect to see the following:

	DNA specific to X	DNA specific to X and Y	DNA specific to Y
Male fetus (XY)	$M + \frac{1}{2} F$	$M + F$	$\frac{1}{2} F$
Female fetus (XX)	$M + F$	$M + F$	0

[0845] In the case where F and M are known, the expected ratios can be computed, and the observed data can be compared to the expected data. In the case where M and F are not known, a threshold can be selected based on historical data. In both cases, the measured amount of DNA at loci specific to both X and Y can be used as a baseline, and the test for the sex of the fetus can be based on the amount of DNA observed on loci specific to only the X chromosome. If that amount is lower than the baseline by an amount roughly equal to $\frac{1}{2} F$, or by an amount that

causes it to fall below a predefined threshold, the fetus is determined to be male, and if that amount is about equal to the baseline, or if is not lower by an amount that causes it to fall below a predefined threshold, the fetus is determined to be female.

[0846] In another embodiment, one can look only at those loci that are common to both the X and the Y chromosomes, often termed the Z chromosome. A subset of the loci on the Z chromosome are typically always A on the X chromosome, and B on the Y chromosome. If SNPs from the Z chromosome are found to have the B genotype, then the fetus is called a male; if the SNPs from the Z chromosome are found to only have A genotype, then the fetus is called a female. In another embodiment, one can look at the loci that are found only on the X chromosome. Contexts such as AA|B are particularly informative as the presence of a B indicates that the fetus has an X chromosome from the father. Contexts such as AB|B are also informative, as we expect to see B present only half as often in the case of a female fetus as compared to a male fetus. In another embodiment, one can look at the SNPs on the Z chromosome where both A and B alleles are present on both the X and the Y chromosome, and where it is known which SNPs are from the paternal Y chromosome, and which are from the paternal X chromosome.

[0847] In an embodiment, it is possible to amplify single nucleotide positions known to varying between the homologous non-recombining (HNR) region shared by chromosome Y and chromosome X. The sequence within this HNR region is largely identical between the X and Y chromosomes. Within this identical region are single nucleotide positions that, while invariant among X chromosomes and among Y chromosomes in the population, are different between the X and Y chromosomes. Each PCR assay could amplify a sequence from loci that are present on both the X and Y chromosomes. Within each amplified sequence would be a single base that can be detected using sequencing or some other method (*see*, for example, U.S. Publication No. 2011/0178719, filed February 3, 2011, which is hereby incorporated by reference in its entirety).

[0848] In an embodiment, the sex of the fetus could be determined from the fetal free floating DNA found in maternal plasma, the method comprising some or all of the following steps: 1) Design PCR (either regular or mini-PCR, plus multiplexing if desired) primers amplify X/Y variant single nucleotide positions within HNR region, 2) obtain maternal plasma, 3) PCR Amplify targets from maternal plasma using HNR X/Y PCR assays, 4) sequence the amplicons, 5) Examine sequence data for presence of Y-allele within one or more of the amplified

sequences. The presence of one or more would indicate a male fetus. Absence of all Y-alleles from all amplicons indicates a female fetus.

[0849] In an embodiment, one could use targeted sequencing to measure the DNA in the maternal plasma and/or the parental genotypes. In an embodiment, one could ignore all sequences that clearly originate from paternally sourced DNA. For example, in the context AA|AB, one could count the number of A sequences and ignore all the B sequences. In order to determine a heterozygosity rate for the above algorithm, one could compare the number of observed A sequences to the expected number of total sequences for the given probe. There are many ways one could calculate an expected number of sequences for each probe on a per sample basis. In an embodiment, it is possible to use historical data to determine what fraction of all sequence reads belongs to each specific probe and then use this empirical fraction, combined with the total number of sequence reads, to estimate the number of sequences at each probe. Another approach could be to target some known homozygous alleles and then use historical data to relate the number of reads at each probe with the number of reads at the known homozygous alleles. For each sample, one could then measure the number of reads at the homozygous alleles and then use this measurement, along with the empirically derived relationships, to estimate the number of sequence reads at each probe.

[0850] In some embodiments, it is possible to determine the sex of the fetus by combining the predictions made by a plurality of methods. In some embodiments the plurality of methods are taken from methods described in this disclosure. In some embodiments, at least one of the plurality of methods are taken from methods described in this disclosure.

[0851] In some embodiments the method described herein can be used to determine the ploidy state of the gestating fetus. In an embodiment, the ploidy calling method uses loci that are specific to the X chromosome, or common to both the X and Y chromosome, but does not make use of any Y-specific loci. In an embodiment, the ploidy calling method uses one or more of the following: loci that are specific to the X chromosome, loci that are common to both the X and Y chromosome, and loci that are specific to the Y chromosome. In an embodiment, where the ratios of sex chromosomes are similar, for example 45,X (Turner Syndrome), 46,XX (normal female) and 47,XXX (trisomy X), the differentiation can be accomplished by comparing the allele distributions to expected allele distributions according to the various hypotheses. In another embodiment, this can be accomplished by comparing the relative number of sequence

reads for the sex chromosomes to one or a plurality of reference chromosomes that are assumed to be euploid. Also note that these methods can be expanded to include aneuploid cases.

[0852] Single Gene Disease Screening

[0853] In an embodiment, a method for determining the ploidy state of the fetus may be extended to enable simultaneous testing for single gene disorders. Single-gene disease diagnosis leverages the same targeted approach used for aneuploidy testing, and requires additional specific targets. In an embodiment, the single gene NPD diagnosis is through linkage analysis. In many cases, direct testing of the cfDNA sample is not reliable, as the presence of maternal DNA makes it virtually impossible to determine if the fetus has inherited the mother's mutation. Detection of a unique paternally-derived allele is less challenging, but is only fully informative if the disease is dominant and carried by the father, limiting the utility of the approach. In an embodiment, the method involves PCR or related amplification approaches.

[0854] In some embodiments, the method involves phasing the abnormal allele with surrounding very tightly linked SNPs in the parents using information from first-degree relatives. Then Parental Support may be run on the targeted sequencing data obtained from these SNPs to determine which homologs, normal or abnormal, were inherited by the fetus from both parents. As long as the SNPs are sufficiently linked, the inheritance of the genotype of the fetus can be determined very reliably. In some embodiments, the method comprises (a) adding a set of SNP loci to densely flank a specified set of common diseases to our multiplex pool for aneuploidy testing; (b) reliably phasing the alleles from these added SNPs with the normal and abnormal alleles based on genetic data from various relatives; and (c) reconstructing the fetal haplotype, or set of phased SNP alleles on the inherited maternal and paternal homologs in the region surrounding the disease locus to determine fetal genotype. In some embodiments additional probes that are closely linked to a disease linked locus are added to the set of polymorphic locus being used for aneuploidy testing.

[0855] Reconstructing fetal diplotype is challenging because the sample is a mixture of maternal and fetal DNA. In some embodiments, the method incorporates relative information to phase the SNPs and disease alleles, then take into account physical distance of the SNPs and recombination data from location specific recombination likelihoods and the data observed from the genetic measurements of the maternal plasma to obtain the most likely genotype of the fetus.

[0856] In an embodiment, a number of additional probes per disease linked locus are included in the set of targeted polymorphic loci; the number of additional probes per disease linked locus may be between 4 and 10, between 11 and 20, between 21 and 40, between 41 and 60, between 61 and 80, or combinations thereof.

[0857] Phasing the diploid data from the parents can be challenging, and there are a number of ways this can be accomplished. Some are discussed in this disclosure, others are described in greater detail in other disclosures (*see, e.g.*, PCT Publ. No. WO2009105531, filed February 9, 2009, and PCT Publ. No. WO2010017214, filed August 4, 2009, which are each hereby incorporated by reference in its entirety). In one embodiment, a parent can be phased by inference by measuring tissue from the parent that is haploid, for example by measuring one or more sperm or eggs. In one embodiment the parent can be phased by inference using the measured genotypic data of a first degree relative such as the parent's parent(s) or siblings. In one embodiment, the parent can be phased by dilution where the DNA is diluted, in one or a plurality of wells, to the point where there is expected to be no more than approximately one copy of each haplotype in each well, and then measuring the DNA in the one or more wells. In one embodiment, the parent genotype can be phased by using computer programs that use population based haplotype frequencies to infer the most likely phase. In one embodiment, the parent can be phased if the phased haplotypic data is known for the other parent, along with the unphased genetic data of one or more genetic offspring of the parents. In some embodiments, the genetic offspring of the parents may be one or more embryos, fetuses, and/or born children. Some of these methods and other methods for phasing one or both parents are disclosed in greater detail in, *e.g.*, U.S. Publ. No. 2011/0033862, filed August 19, 2010; U.S. Publ. No. 2011/0178719, filed Feb. 3, 2011; U.S. Publ. No. 2007/0184467, filed Nov. 22, 2006; U.S. Publ. No. 2008/0243398, filed March 17, 2008, which are each hereby incorporated by reference in its entirety.

[0858] *Fetal Genome Reconstruction*

[0859] In one aspect, the invention features methods for determining a haplotype of a fetus. In various embodiments, this method allows one to determine which polymorphic loci (such as SNPs) were inherited by the fetus and to reconstruct which homologs (including recombination events) are present in the fetus (and thereby interpolate the sequence between the polymorphic

loci). If desired, essentially the entire genome of the fetus can be reconstructed. If there is some remaining ambiguity in the genome of the fetus (such as in intervals with a crossover), this ambiguity can be minimized if desired by analyzing additional polymorphic loci. In various embodiments, the polymorphic loci are chosen to cover one or more of the chromosomes at a density to reduce any ambiguity to a desired level. This method has important applications for the detection of polymorphisms or other mutations of interest in a fetus since it enables their detection based on linkage (such as the presence of linked polymorphic loci in the fetal genome) rather than by directly detecting the polymorphism or other mutation of interest in the fetal genome. For example, if a parent is a carrier for a mutation associated with cystic fibrosis (CF), a nucleic acid sample that includes maternal DNA from the mother of the fetus and fetal DNA from the fetus can be analyzed to determine whether the fetal DNA include the haplotype containing the CF mutation. In particular, polymorphic loci can be analyzed to determine whether the fetal DNA includes the haplotype containing the CF mutation without having to detect the CF mutation itself in the fetal DNA. This is useful in screening for one or more mutations, such as disease-linked mutations, without having to directly detect the mutations.

[0860] In some embodiments, the method involves determining a parental haplotype (*e.g.*, a haplotype of the mother or father of the fetus). In some embodiments, this determination is made without using data from a relative of the mother or father. In some embodiments, a parental haplotype is determined using a dilution approach followed by SNP genotyping or sequencing as described herein and elsewhere (*see, e.g.*, U.S. Publ. No. 2011/0033862, filed August 19, 2010, which is hereby incorporated by reference in its entirety). Because the DNA is diluted, it is unlikely that more than one haplotype is in the same fraction (or tube). Thus, there may be effectively a single molecule of DNA in the tube, which allows the haplotype on a single DNA molecule to be determined. In some embodiments, the method includes dividing a DNA sample into a plurality of fractions such that at least one of the fractions includes one chromosome or one chromosome segment from a pair of chromosomes, and genotyping (*e.g.*, determining the presence of two or more polymorphic loci) the DNA sample in at least one of the fractions, thereby determining a parental haplotype. In some embodiments, the genotyping involves sequencing (such as shotgun sequencing). In some embodiments, the genotyping involves use of a SNP array to detect polymorphic loci, such as at least 25; 50; 75; 100; 300; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 15,000; 19,000; 20,000; 25,000; 27,000; 28,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci. In some embodiments, the genotyping

involves the use of multiplex PCR. In some embodiments, the method involves contacting the sample in a fraction with a library of primers that simultaneously hybridize to at least 25; 50; 75; 100; 300; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 15,000; 19,000; 20,000; 25,000; 27,000; 28,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci (such as SNPs) to produce a reaction mixture; and subjecting the reaction mixture to primer extension reaction conditions to produce amplified products that are measured with a high throughput sequencer to produce sequencing data.

[0861] In some embodiments, a haplotype of the mother is determined by any of the methods described herein using data from a relative of the mother. In some embodiments, a haplotype of the father is determined by any of the methods described herein using data from a relative of the father. In some embodiments, a haplotype is determined for both the father and the mother. In some embodiments, a SNP array is used to determine the presence of at least 25; 50; 75; 100; 300; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 15,000; 19,000; 20,000; 25,000; 27,000; 28,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci in a DNA sample from the mother (or father) and a relative of the mother (or father). In some embodiments, the method involves contacting a DNA sample from the mother (or father) and/or a relative of the mother (or father) with a library of primers that simultaneously hybridize to at least 25; 50; 75; 100; 300; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 15,000; 19,000; 20,000; 25,000; 27,000; 28,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci (such as SNPs) to produce a reaction mixture; and subjecting the reaction mixture to primer extension reaction conditions to produce amplified products that are measured with a high throughput sequencer to produce sequencing data. The parental haplotype may be determined based on the SNP array or sequencing data. In some embodiments, parental data may be phased by methods described or referred to elsewhere in this document.

[0862] This parental haplotype data can be used to determine if the fetus inherited the parental haplotype. In some embodiments, a nucleic acid sample that includes maternal DNA from the mother of the fetus and fetal DNA from the fetus is analyzed using a SNP array to detect at least 25; 50; 75; 100; 300; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 15,000; 19,000; 20,000; 25,000; 27,000; 28,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci. In some embodiments, a nucleic acid sample that includes maternal DNA from the mother of the fetus and fetal DNA from the fetus is analyzed by contacting the sample with a library of primers that simultaneously hybridize to at least 25; 50; 75; 100; 300; 500; 750; 1,000; 2,000;

5,000; 7,500; 10,000; 15,000; 19,000; 20,000; 25,000; 27,000; 28,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci (such as SNPs) to produce a reaction mixture. In some embodiments, the reaction mixture is subjected to primer extension reaction conditions to produce amplified products. In some embodiments, the amplified products are measured with a high throughput sequencer to produce sequencing data. In various embodiments, the SNP array or sequencing data is used to determine a parental haplotype by using data about the probability of chromosomes crossing over at different locations in a chromosome (such as by using recombination data such as may be found in the HapMap database to create a recombination risk score for any interval) to model dependence between polymorphic alleles on the chromosome. In some embodiments, allele counts at the polymorphic loci are calculated on a computer based on the sequencing data. In some embodiments, a plurality of ploidy hypotheses each pertaining to a different possible ploidy state of the chromosome are created on a computer; a model (such as a joint distribution model) for the expected allele counts at the polymorphic loci on the chromosome is built on a computer for each ploidy hypothesis; a relative probability of each of the ploidy hypotheses is determined on a computer using the joint distribution model and the allele counts; and the ploidy state of the fetus is called by selecting the ploidy state corresponding to the hypothesis with the greatest probability. In some embodiments, building a joint distribution model for allele counts and the step of determining the relative probability of each hypothesis are done using a method that does not require the use of a reference chromosome.

[0863] In some embodiments, a fetal haplotype is determined for one or more chromosomes taken from the group consisting of chromosomes 13, 18, 21, X, and Y. In some embodiments, a fetal haplotype is determined for all of the fetal chromosomes. In various embodiments, the method determines essentially the entire genome of the fetus. In some embodiments, the haplotype is determined for at least 30, 40, 50, 60, 70, 80, 90, or 95% of the genome of the fetus. In some embodiments, the haplotype determination of the fetus includes information about which allele is present for at least 25; 50; 75; 100; 300; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 15,000; 19,000; 20,000; 25,000; 27,000; 28,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different polymorphic loci.

[0864] *Compositions of DNA*

[0865] When performing an informatics analysis on sequencing data measured on a mixture of fetal and maternal blood to determine genomic information pertaining to the fetus, for example the ploidy state of the fetus, it may be advantageous to measure the allele distributions at a set of alleles. Unfortunately, in many cases, such as when attempting to determine the ploidy state of a fetus from the DNA mixture found in the plasma of a maternal blood sample, the amount of DNA available is not sufficient to directly measure the allele distributions with good fidelity in the mixture. In these cases, amplification of the DNA mixture will provide sufficient numbers of DNA molecules that the desired allele distributions may be measured with good fidelity. However, current methods of amplification typically used in the amplification of DNA for sequencing are often very biased, meaning that they do not amplify both alleles at a polymorphic locus by the same amount. A biased amplification can result in allele distributions that are quite different from the allele distributions in the original mixture. For most purposes, highly accurate measurements of the relative amounts of alleles present at polymorphic loci are not needed. In contrast, in an embodiment of the present disclosure, amplification or enrichment methods that specifically enrich polymorphic alleles and preserve allelic ratios is advantageous.

[0866] A number of methods are described herein that may be used to preferentially enrich a sample of DNA at a plurality of loci in a way that minimizes allelic bias. Some examples are using circularizing probes to target a plurality of loci where the 3' ends and 5' ends of the pre-circularized probe are designed to hybridize to bases that are one or a few positions away from the polymorphic sites of the targeted allele. Another is to use PCR probes where the 3' end PCR probe is designed to hybridize to bases that are one or a few positions away from the polymorphic sites of the targeted allele. Another is to use a split and pool approach to create mixtures of DNA where the preferentially enriched loci are enriched with low allelic bias without the drawbacks of direct multiplexing. Another is to use a hybrid capture approach where the capture probes are designed such that the region of the capture probe that is designed to hybridize to the DNA flanking the polymorphic site of the target is separated from the polymorphic site by one or a small number of bases.

[0867] In the case where measured allele distributions at a set of polymorphic loci are used to determine the ploidy state of an individual, it is desirable to preserve the relative amounts of alleles in a sample of DNA as it is prepared for genetic measurements. This preparation may involve WGA amplification, targeted amplification, selective enrichment techniques, hybrid

capture techniques, circularizing probes or other methods meant to amplify the amount of DNA and/or selectively enhance the presence of molecules of DNA that correspond to certain alleles.

[0868] In some embodiments of the present disclosure, there is a set of DNA probes designed to target loci where the loci have maximal minor allele frequencies. In some embodiments of the present disclosure, there is a set of probes that are designed to target where the loci have the maximum likelihood of the fetus having a highly informative SNP at those loci. In some embodiments of the present disclosure, there is a set of probes that are designed to target loci where the probes are optimized for a given population subgroup. In some embodiments of the present disclosure, there is a set of probes that are designed to target loci where the probes are optimized for a given mix of population subgroups. In some embodiments of the present disclosure, there is a set of probes that are designed to target loci where the probes are optimized for a given pair of parents which are from different population subgroups that have different minor allele frequency profiles. In some embodiments of the present disclosure, there is a circularized strand of DNA that comprises at least one base pair that annealed to a piece of DNA that is of fetal origin. In some embodiments of the present disclosure, there is a circularized strand of DNA that comprises at least one base pair that annealed to a piece of DNA that is of placental origin. In some embodiments of the present disclosure, there is a circularized strand of DNA that circularized while at least some of the nucleotides were annealed to DNA that was of fetal origin. In some embodiments of the present disclosure, there is a circularized strand of DNA that circularized while at least some of the nucleotides were annealed to DNA that was of placental origin. In some embodiments of the present disclosure, there is a set of probes wherein some of the probes target single tandem repeats, and some of the probes target single nucleotide polymorphisms. In some embodiments, the loci are selected for the purpose of non-invasive prenatal diagnosis. In some embodiments, the probes are used for the purpose of non-invasive prenatal diagnosis. In some embodiments, the loci are targeted using a method that could include circularizing probes, MIPs, capture by hybridization probes, probes on a SNP array, or combinations thereof. In some embodiments, the probes are used as circularizing probes, MIPs, capture by hybridization probes, probes on a SNP array, or combinations thereof. In some embodiments, the loci are sequenced for the purpose of non-invasive prenatal diagnosis.

[0869] In the case where the relative informativeness of a sequence is greater when combined with relevant parent contexts, it follows that maximizing the number of sequence reads that contain a SNP for which the parental context is known may maximize the informativeness of

the set of sequencing reads on the mixed sample. In an embodiment, the number of sequence reads that contain a SNP for which the parent contexts are known may be enhanced by using qPCR to preferentially amplify specific sequences. In an embodiment, the number of sequence reads that contain a SNP for which the parent contexts are known may be enhanced by using circularizing probes (for example, MIPs) to preferentially amplify specific sequences. In an embodiment, the number of sequence reads that contain a SNP for which the parent contexts are known may be enhanced by using a capture by hybridization method (for example SURESELECT) to preferentially amplify specific sequences. Different methods may be used to enhance the number of sequence reads that contain a SNP for which the parent contexts are known. In an embodiment, the targeting may be accomplished by extension ligation, ligation without extension, capture by hybridization, or PCR.

[0870] In a sample of fragmented genomic DNA, a fraction of the DNA sequences map uniquely to individual chromosomes; other DNA sequences may be found on different chromosomes. Note that DNA found in plasma, whether maternal or fetal in origin is typically fragmented, often at lengths under 500 bp. In a typical genomic sample, roughly 3.3% of the mappable sequences will map to chromosome 13; 2.2% of the mappable sequences will map to chromosome 18; 1.35% of the mappable sequences will map to chromosome 21; 4.5% of the mappable sequences will map to chromosome X in a female; 2.25% of the mappable sequences will map to chromosome X (in a male); and 0.73% of the mappable sequences will map to chromosome Y (in a male). These are the chromosomes that are most likely to be aneuploid in a fetus. Also, among short sequences, approximately 1 in 20 sequences will contain a SNP, using the SNPs contained on dbSNP. The proportion may well be higher given that there may be many SNPs that have not been discovered.

[0871] In an embodiment of the present disclosure, targeting methods may be used to enhance the fraction of DNA in a sample of DNA that map to a given chromosome such that the fraction significantly exceeds the percentages listed above that are typical for genomic samples. In an embodiment of the present disclosure, targeting methods may be used to enhance the fraction of DNA in a sample of DNA such that the percentage of sequences that contain a SNP are significantly greater than what may be found in typical for genomic samples. In an embodiment of the present disclosure, targeting methods may be used to target DNA from a chromosome or from a set of SNPs in a mixture of maternal and fetal DNA for the purposes of prenatal diagnosis.

[0872] Note that a method has been reported (U.S. Patent 7,888,017) for determining fetal aneuploidy by counting the number of reads that map to a suspect chromosome and comparing it to the number of reads that map to a reference chromosome, and using the assumption that an overabundance of reads on the suspect chromosome corresponds to a triploidy in the fetus at that chromosome. Those methods for prenatal diagnosis would not make use of targeting of any sort, nor do they describe the use of targeting for prenatal diagnosis.

[0873] By making use of targeting approaches in sequencing the mixed sample, it may be possible to achieve a certain level of accuracy with fewer sequence reads. The accuracy may refer to sensitivity, it may refer to specificity, or it may refer to some combination thereof. The desired level of accuracy may be between 90% and 95%; it may be between 95% and 98%; it may be between 98% and 99%; it may be between 99% and 99.5%; it may be between 99.5% and 99.9%; it may be between 99.9% and 99.99%; it may be between 99.99% and 99.999%, it may be between 99.999% and 100%. Levels of accuracy above 95% may be referred to as high accuracy.

[0874] There are a number of published methods in the prior art that demonstrate how one may determine the ploidy state of a fetus from a mixed sample of maternal and fetal DNA, for example: G.J. W. Liao et al. Clinical Chemistry 2011; 57(1) pp. 92-101. These methods focus on thousands of locations along each chromosome. The number of locations along a chromosome that may be targeted while still resulting in a high accuracy ploidy determination on a fetus, for a given number of sequence reads, from a mixed sample of DNA is unexpectedly low. In an embodiment of the present disclosure, an accurate ploidy determination may be made by using targeted sequencing, using any method of targeting, for example qPCR, ligand mediated PCR, other PCR methods, capture by hybridization, or circularizing probes, wherein the number of loci along a chromosome that need to be targeted may be between 5,000 and 2,000 loci; it may be between 2,000 and 1,000 loci; it may be between 1,000 and 500 loci; it may be between 500 and 300 loci; it may be between 300 and 200 loci; it may be between 200 and 150 loci; it may be between 150 and 100 loci; it may be between 100 and 50 loci; it may be between 50 and 20 loci; it may be between 20 and 10 loci. Optimally, it may be between 100 and 500 loci. The high level of accuracy may be achieved by targeting a small number of loci and executing an unexpectedly small number of sequence reads. The number of reads may be between 100 million and 50 million reads; the number of reads may be between 50 million and 20 million reads; the number of reads may be between 20 million and 10 million reads; the

number of reads may be between 10 million and 5 million reads; the number of reads may be between 5 million and 2 million reads; the number of reads may be between 2 million and 1 million; the number of reads may be between 1 million and 500,000; the number of reads may be between 500,000 and 200,000; the number of reads may be between 200,000 and 100,000; the number of reads may be between 100,000 and 50,000; the number of reads may be between 50,000 and 20,000; the number of reads may be between 20,000 and 10,000; the number of reads may be below 10,000. Fewer number of read are necessary for larger amounts of input DNA.

[0875] In some embodiments, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to chromosome 13 is greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, greater than 20%, greater than 25%, or greater than 30%. In some embodiments of the present disclosure, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to chromosome 18 is greater than 3%, greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, greater than 20%, greater than 25%, or greater than 30%. In some embodiments of the present disclosure, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to chromosome 21 is greater than 2%, greater than 3%, greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, greater than 20%, greater than 25%, or greater than 30%. In some embodiments of the present disclosure, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to chromosome X is greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, greater than 20%, greater than 25%, or greater than 30%. In some embodiments of the present disclosure, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to chromosome Y is greater than 1%, greater than 2%, greater than 3%, greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, greater than 20%, greater than 25%, or greater than 30%.

[0876] In some embodiments, a composition is described comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to a chromosome, and that contains at least one single nucleotide polymorphism is greater than 0.2%, greater than 0.3%, greater than 0.4%, greater than 0.5%, greater than 0.6%, greater than 0.7%, greater than 0.8%, greater than 0.9%, greater than 1%, greater than 1.2%, greater than 1.4%, greater than 1.6%, greater than 1.8%, greater than 2%, greater than 2.5%, greater than 3%, greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, or greater than 20%, and where the chromosome is taken from the group 13, 18, 21, X, or Y. In some embodiments of the present disclosure, there is a composition comprising a mixture of DNA of fetal origin, and DNA of maternal origin, wherein the percent of sequences that uniquely map to a chromosome and that contain at least one single nucleotide polymorphism from a set of single nucleotide polymorphisms is greater than 0.15%, greater than 0.2%, greater than 0.3%, greater than 0.4%, greater than 0.5%, greater than 0.6%, greater than 0.7%, greater than 0.8%, greater than 0.9%, greater than 1%, greater than 1.2%, greater than 1.4%, greater than 1.6%, greater than 1.8%, greater than 2%, greater than 2.5%, greater than 3%, greater than 4%, greater than 5%, greater than 6%, greater than 7%, greater than 8%, greater than 9%, greater than 10%, greater than 12%, greater than 15%, or greater than 20%, where the chromosome is taken from the set of chromosome 13, 18, 21, X and Y, and where the number of single nucleotide polymorphisms in the set of single nucleotide polymorphisms is between 1 and 10, between 10 and 20, between 20 and 50, between 50 and 100, between 100 and 200, between 200 and 500, between 500 and 1,000, between 1,000 and 2,000, between 2,000 and 5,000, between 5,000 and 10,000, between 10,000 and 20,000, between 20,000 and 50,000, and between 50,000 and 100,000.

[0877] In theory, each cycle in the amplification doubles the amount of DNA present; however, in reality, the degree of amplification is slightly lower than two. In theory, amplification, including targeted amplification, will result in bias free amplification of a DNA mixture; in reality, however, different alleles tend to be amplified to a different extent than other alleles. When DNA is amplified, the degree of allelic bias typically increases with the number of amplification steps. In some embodiments, the methods described herein involve amplifying DNA with a low level of allelic bias. Since the allelic bias compounds with each additional cycle, one can determine the per cycle allelic bias by calculating the n th root of the overall bias where n is the base 2 logarithm of degree of enrichment. In some embodiments, there is a composition

comprising a second mixture of DNA, where the second mixture of DNA has been preferentially enriched at a plurality of polymorphic loci from a first mixture of DNA where the degree of enrichment is at least 10, at least 100, at least 1,000, at least 10,000, at least 100,000 or at least 1,000,000, and where the ratio of the alleles in the second mixture of DNA at each locus differs from the ratio of the alleles at that locus in the first mixture of DNA by a factor that is, on average, less than 1,000%, 500%, 200%, 100%, 50%, 20%, 10%, 5%, 2%, 1%, 0.5%, 0.2%, 0.1%, 0.05%, 0.02%, or 0.01%. In some embodiments, there is a composition comprising a second mixture of DNA, where the second mixture of DNA has been preferentially enriched at a plurality of polymorphic loci from a first mixture of DNA where the per cycle allelic bias for the plurality of polymorphic loci is, on average, less than 10%, 5%, 2%, 1%, 0.5%, 0.2%, 0.1%, 0.05%, or 0.02%. In some embodiments, the plurality of polymorphic loci comprises at least 10 loci, at least 20 loci, at least 50 loci, at least 100 loci, at least 200 loci, at least 500 loci, at least 1,000 loci, at least 2,000 loci, at least 5,000 loci, at least 10,000 loci, at least 20,000 loci, or at least 50,000 loci.

[0878] Some Embodiments

[0879] In some embodiments, a method is disclosed herein for generating a report disclosing the determined ploidy status of a chromosome in a gestating fetus, the method comprising: obtaining a first sample that contains DNA from the mother of the fetus and DNA from the fetus; obtaining genotypic data from one or both parents of the fetus; preparing the first sample by isolating the DNA so as to obtain a prepared sample; measuring the DNA in the prepared sample at a plurality of polymorphic loci; calculating, on a computer, allele counts or allele count probabilities at the plurality of polymorphic loci from the DNA measurements made on the prepared sample; creating, on a computer, a plurality of ploidy hypotheses concerning expected allele count probabilities at the plurality of polymorphic loci on the chromosome for different possible ploidy states of the chromosome; building, on a computer, a joint distribution model for allele count probability of each polymorphic locus on the chromosome for each ploidy hypothesis using genotypic data from the one or both parents of the fetus; determining, on a computer, a relative probability of each of the ploidy hypotheses using the joint distribution model and the allele count probabilities calculated for the prepared sample; calling the ploidy state of the fetus by

selecting the ploidy state corresponding to the hypothesis with the greatest probability; and generating a report disclosing the determined ploidy status.

[0880] In some embodiments, the method is used to determine the ploidy state of a plurality of gestating fetuses in a plurality of respective mothers, the method further comprising: determining the percent of DNA that is of fetal origin in each of the prepared samples; and wherein the step of measuring the DNA in the prepared sample is done by sequencing a number of DNA molecules in each of the prepared samples, where more molecules of DNA are sequenced from those prepared samples that have a smaller fraction of fetal DNA than those prepared samples that have a larger fraction of fetal DNA.

[0881] In some embodiments, the method is used to determine the ploidy state of a plurality of gestating fetuses in a plurality of respective mothers, and where the measuring the DNA in the prepared sample is done, for each of the fetuses, by sequencing a first fraction of the prepared sample of DNA to give a first set of measurements, the method further comprising: making a first relative probability determination for each of the ploidy hypotheses for each of the fetuses, given the first set of DNA measurements; resequencing a second fraction of the prepared sample from those fetuses where the first relative probability determination for each of the ploidy hypotheses indicates that a ploidy hypothesis corresponding to an aneuploid fetus has a significant but not conclusive probability, to give a second set of measurements; making a second relative probability determination for ploidy hypotheses for the fetuses using the second set of measurements and optionally also the first set of measurements; and calling the ploidy states of the fetuses whose second sample was resequenced by selecting the ploidy state corresponding to the hypothesis with the greatest probability as determined by the second relative probability determination.

[0882] In some embodiments, a composition of matter is disclosed, the composition of matter comprising: a sample of preferentially enriched DNA, wherein the sample of preferentially enriched DNA has been preferentially enriched at a plurality of polymorphic loci from a first sample of DNA, wherein the first sample of DNA consisted of a mixture of maternal DNA and fetal DNA derived from maternal plasma, where the degree of enrichment is at least a factor of 2, and wherein the allelic bias between the first sample and the preferentially enriched sample is, on average, selected from the group consisting of less than 2%, less than 1%, less than 0.5%, less than 0.2%, less than 0.1%, less than 0.05%, less than 0.02%, and less than 0.01%. In some embodiments, a method is disclosed to create a sample of such preferentially enriched DNA.

[0883] In some embodiment, a method is disclosed for determining the presence or absence of a fetal aneuploidy in a maternal tissue sample comprising fetal and maternal genomic DNA, wherein the method comprises: (a) obtaining a mixture of fetal and maternal genomic DNA from said maternal tissue sample; (b) selectively enriching the mixture of fetal and maternal DNA at a plurality of polymorphic alleles; (c) distributing selectively enriched fragments from the mixture of fetal and maternal genomic DNA of step a to provide reaction samples comprising a single genomic DNA molecule or amplification products of a single genomic DNA molecule; (d) conducting massively parallel DNA sequencing of the selectively enriched fragments of genomic DNA in the reaction samples of step c) to determine the sequence of said selectively enriched fragments; (e) identifying the chromosomes to which the sequences obtained in step d) belong; (f) analyzing the data of step d) to determine i) the number of fragments of genomic DNA from step d) that belong to at least one first target chromosome that is presumed to be diploid in both the mother and the fetus, and ii) the number of fragments of genomic DNA from step d) that belong to a second target chromosome, wherein said second chromosome is suspected to be aneuploid in the fetus; (g) calculating an expected distribution of the number of fragments of genomic DNA from step d) for the second target chromosome if the second target chromosome is euploid, using the number determined in step f) part i); (h) calculating an expected distribution of the number of fragments of genomic DNA from step d) for the second target chromosome if the second target chromosome is aneuploid, using the first number is step f) part i) and an estimated fraction of fetal DNA found in the mixture of step b); and (i) using a maximum likelihood or maximum a posteriori approach to determine whether the number of fragments of genomic DNA determined in step f) part ii) is more likely to be part of the distribution calculated in step g) or the distribution calculated in step h); thereby indicating the presence or absence of a fetal aneuploidy.

[0884] Exemplary Cancer Diagnostic Methods

[0885] Note that it has been demonstrated that DNA that originated from cancer that is living in a host can be found in the blood of the host. In the same way that genetic diagnoses can be made from the measurement of mixed DNA found in maternal blood, genetic diagnoses can equally well be made from the measurement of mixed DNA found in host blood. The genetic diagnoses may include aneuploidy states, or gene mutations. Any claim in the instant disclosure that reads

on determining the ploidy state or genetic state of a fetus from the measurements made on maternal blood can equally well read on determining the ploidy state or genetic state of a cancer from the measurements on host blood.

[0886] In some embodiments, a method of the present disclosure allows one to determine the ploidy status of a cancer, the method including obtaining a mixed sample that contains genetic material from the host, and genetic material from the cancer; measuring the DNA in the mixed sample; calculating the fraction of DNA that is of cancer origin in the mixed sample; and determining the ploidy status of the cancer using the measurements made on the mixed sample and the calculated fraction. In some embodiments, the method may further include administering a cancer therapeutic based on the determination of the ploidy state of the cancer. In some embodiments, the method may further include administering a cancer therapeutic based on the determination of the ploidy state of the cancer, wherein the cancer therapeutic is taken from the group comprising a pharmaceutical, a biologic therapeutic, and antibody based therapy and combination thereof.

[0887] Exemplary Clinical Actions

[0888] In some embodiments, any of the methods include taking a clinical action based on a result of a method of the invention (such as the determination of the presence or absence of a polymorphism or mutation, ploidy state, or paternity). In some embodiments in which an embryo or fetus has one or more one or more polymorphisms or mutations of interest (such as a CNV) based on a result of a method of the invention, the clinical action includes performing additional testing (such as testing to confirm the presence of the polymorphism or mutation), not implanting the embryo for IVF, implanting a different embryo for IVF, terminating a pregnancy, preparing for a special needs child, or undergoing an intervention designed to decrease the severity of the phenotypic presentation of a genetic disorder. In some embodiments, the clinical action is selected from the group consisting of performing an ultrasound, amniocentesis on the fetus, amniocentesis on a subsequent fetus that inherits genetic material from the mother and/or father, chorion villus biopsy on the fetus, chorion villus biopsy on a subsequent fetus that inherits genetic material from the mother and/or father, *in vitro* fertilization, preimplantation genetic diagnosis on one or more embryos that inherited genetic material from the mother and/or father, karyotyping on the mother, karyotyping on the father, fetal echocardiogram (such as an

echocardiogram of a fetus with trisomy 21, 18, or 13, monosomy X, or a microdeletion), and combinations thereof. In some embodiments, the clinical action is selected from the group consisting of administering growth hormone to a born child with monosomy X (such as administration starting at ~9 months), administering calcium to a born child with a 22q deletion (such as DiGeorge syndrome), administering an androgen such as testosterone to a born child with 47,XXY (such as one injection per month for 3 months of 25 mg testosterone enanthate to an infant or toddler), performing a test for cancer on a woman with a complete or partial molar pregnancy (such as a triploid fetus), administering a therapy for cancer such as a chemotherapeutic agent to a woman with a complete or partial molar pregnancy (such as a triploid fetus), screening a fetus determined to be male (such as a fetus determined to be male using a method of the invention) for one or more X-linked genetic disorders such as Duchenne muscular dystrophy (DMD), adrenoleukodystrophy, or hemophilia, performing amniocentesis on a male fetus at risk for an X-linked disorder, administering dexamethasone to a women with a female fetus at risk male (such as a fetus determined to be female using a method of the invention) for congenital adrenal hyperplasia, performing amniocentesis on a female fetus at risk for congenital adrenal hyperplasia, administering killed vaccines (instead of live vaccines) or not administering certain vaccines to a born child who is (or is suspected of being) immune deficient from a 22q11.2 deletion, performing occupational and/or physical therapy, performing early intervention in education, delivering the baby at a tertiary care center with a NICU and/or having pediatric specialists available at delivery, behavioral intervention for born child (such as a child with XXX, XXY, or XYY), and combinations thereof.

[0889] In some embodiments, ultrasound or another screening test is performed on a women determined to have multiple pregnancies (such as twins) to determine whether or not two or more of the fetus are monochorionic. Monozygotic twins result from ovulation and fertilization of a single oocyte, with subsequent division of the zygote; placentation may be dichorionic or monochorionic. Dizygotic twins occur from ovulation and fertilization of two oocytes, which usually results in dichorionic placentation. Monochorionic twins have a risk of twin-to-twin transfusion syndrome, which may cause unequal distribution of blood between fetuses that results in differences in their growth and development, sometimes resulting in stillbirth. Thus, twins determined to be monozygotic twins using a method of the invention are desirably tested (such as by ultrasound) to determine if they are monochorionic twins, and if so, these twins can

be monitored (such as bi-weekly ultrasounds from 16 weeks) for signs of win-to-twin transfusion syndrome.

[0890] In some embodiments in which an embryo or fetus does not have one or more one or more polymorphisms or mutations of interest (such as a CNV) based on a result of a method of the invention, the clinical action includes implanting the embryo for IVF or continuing a pregnancy. In some embodiments, the clinical action is additional testing to confirm the absence of the polymorphism or mutation selected from the group consisting of performing an ultrasound, amniocentesis, chorion villus biopsy, and combinations thereof.

[0891] In some embodiments in which an individual has one or more polymorphisms or mutations (such as a polymorphism or mutation associated with a disease or disorder such as cancer or an increased risk for a disease or disorder such as cancer) based on a result of a method of the invention, the clinical action includes performing additional testing or administering one or more therapies for a disease or disorder (such as a therapy for cancer, a therapy for the specific type of cancer or type of mutation the individual is diagnosed with, or any of the therapies disclosed herein). In some embodiments, the clinical action is additional testing to confirm the presence or absence of a polymorphism or mutation selected from the group consisting of biopsy, surgery, medical imaging (such as a mammogram or an ultrasound), and combinations thereof.

[0892] In some embodiments, the additional testing includes performing the same or a different method (such as any of the methods described herein) to confirm the presence or absence of the polymorphism or mutation (such as a CNV), such as testing either a second fraction of the same sample that was tested or a different sample from the same individual (such as the same pregnant mother, fetus, embryo, or individual at increased risk for cancer). In some embodiments, the additional testing is performed for an individual for whom the probability of a polymorphism or mutation (such as a CNV) is above a threshold value. In some embodiments, the additional testing is performed for an individual for whom the confidence or z-score for the determination of a polymorphism or mutation (such as a CNV) is above a threshold value (such as additional testing to confirm the presence of a likely polymorphism or mutation). In some embodiments, the additional testing is performed for an individual for whom the confidence or z-score for the determination of a polymorphism or mutation (such as a CNV) is between minimum and maximum threshold values (such as additional testing to increase the confidence that the initial result is correct). In some embodiments, the additional testing is performed for an individual for whom the confidence for the determination of the presence or absence of a polymorphism or

mutation (such as a CNV) is below a threshold value (such as a “no call” result due to not being able to determine the presence or absence of the CNV with sufficient confidence). An exemplary Z core is calculated in Chiu et al. BMJ 2011;342:c7401 (which is hereby incorporated by reference in its entirety) in which chromosome 21 is used as an example and can be replaced with any other chromosome or chromosome segment in the test sample.

[0893]

Z

score for percentage chromosome 21 in test case = ((percentage chromosome 21 in test case) – (mean percentage chromosome 21 in reference controls)) / (standard deviation of percentage chromosome 21 in reference controls).

[0894] In some embodiments, the additional testing is performed for an individual for whom the initial sample did not meet quality control guidelines or had a fetal fraction or a tumor fraction below a threshold value. In some embodiments, the method includes selecting an individual for additional testing based on the result of a method of the invention, the probability of the result, the confidence of the result, or the z-score; and performing the additional testing on the individual (such as on the same or a different sample). In some embodiments, a subject diagnosed with a disease or disorder (such as cancer) undergoes repeat testing using a method of the invention or known testing for the disease or disorder at multiple time points to monitor the progression of the disease or disorder or the remission or reoccurrence of the disease or disorder.

[0895] Exemplary Implementation Methods

[0896] Any of the embodiments disclosed herein may be implemented in digital electronic circuitry, integrated circuitry, specially designed ASICs (application-specific integrated circuits), computer hardware, firmware, software, or in combinations thereof. Apparatus of the presently disclosed embodiments can be implemented in a computer program product tangibly embodied in a machine-readable storage device for execution by a programmable processor; and method steps of the presently disclosed embodiments can be performed by a programmable processor executing a program of instructions to perform functions of the presently disclosed embodiments by operating on input data and generating output. The presently disclosed embodiments can be implemented advantageously in one or more computer programs that are

executable and/or interpretable on a programmable system including at least one programmable processor, which may be special or general purpose, coupled to receive data and instructions from, and to transmit data and instructions to, a storage system, at least one input device, and at least one output device. Each computer program can be implemented in a high-level procedural or object-oriented programming language or in assembly or machine language if desired; and in any case, the language can be a compiled or interpreted language. A computer program may be deployed in any form, including as a stand-alone program, or as a module, component, subroutine, or other unit suitable for use in a computing environment. A computer program may be deployed to be executed or interpreted on one computer or on multiple computers at one site, or distributed across multiple sites and interconnected by a communication network.

[0897] Computer readable storage media, as used herein, refers to physical or tangible storage (as opposed to signals) and includes without limitation volatile and non-volatile, removable and non-removable media implemented in any method or technology for the tangible storage of information such as computer-readable instructions, data structures, program modules or other data. Computer readable storage media includes, but is not limited to, RAM, ROM, EPROM, EEPROM, flash memory or other solid state memory technology, CD-ROM, DVD, or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other physical or material medium which can be used to tangibly store the desired information or data or instructions and which can be accessed by a computer or processor.

[0898] In some embodiments, the invention features a computer configured to accomplish one or more of the *in vitro* methods described herein. In some embodiments, the data is analyzed by the computer system as described herein. In some embodiments, genetic data (such as sequencing or microarray data) from at least 100; 200; 500; 750; 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different loci is analyzed by the computer is less than 200, 100, 60, 30, 20, 10, 5, or 1 minute, or in less than 30 or 10 seconds to detect the present or absence of a mutation (such as a CNV or SNV) at the loci.

[0899] Any of the methods described herein may include the output of data in a physical format, such as on a computer screen, or on a paper printout. In explanations of any embodiments elsewhere in this document, it should be understood that the described methods may be combined with the output of the actionable data in a format that can be acted upon by a physician. In addition, the described methods may be combined with the actual execution of a clinical

decision that results in a clinical treatment, or the execution of a clinical decision to make no action. Some of the embodiments described in the document for determining genetic data pertaining to a target individual may be combined with the decision to select one or more embryos for transfer in the context of IVF, optionally combined with the process of transferring the embryo to the womb of the prospective mother. Some of the embodiments described in the document for determining genetic data pertaining to a target individual may be combined with the notification of a potential chromosomal abnormality, or lack thereof, with a medical professional, optionally combined with the decision to abort, or to not abort, a fetus in the context of prenatal diagnosis. Some of the embodiments described herein may be combined with the output of the actionable data, and the execution of a clinical decision that results in a clinical treatment, or the execution of a clinical decision to make no action.

[0900] Exemplary Diagnostic Boxes

[0901] In an embodiment, the present disclosure comprises a diagnostic box that is capable of partly or completely carrying out any of the methods described in this disclosure. In an embodiment, the diagnostic box may be located at a physician's office, a hospital laboratory, or any suitable location reasonably proximal to the point of patient care. The box may be able to run the entire method in a wholly automated fashion, or the box may require one or a number of steps to be completed manually by a technician. In an embodiment, the box may be able to analyze at least the genotypic data measured on the maternal plasma. In an embodiment, the box may be linked to means to transmit the genotypic data measured on the diagnostic box to an external computation facility which may then analyze the genotypic data, and possibly also generate a report. The diagnostic box may include a robotic unit that is capable of transferring aqueous or liquid samples from one container to another. It may comprise a number of reagents, both solid and liquid. It may comprise a high throughput sequencer. It may comprise a computer.

[0902] Experimental Section

[0903] The presently disclosed embodiments are described in the following Examples, which are set forth to aid in the understanding of the disclosure, and should not be construed to limit in any way the scope of the disclosure as defined in the claims which follow thereafter. The

following examples are put forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to use the described embodiments, and are not intended to limit the scope of the disclosure nor are they intended to represent that the Examples below are all or the only experiments performed. Efforts have been made to ensure accuracy with respect to numbers used (*e.g.* amounts, temperature, etc.) but some experimental errors and deviations should be accounted for. Unless indicated otherwise, parts are parts by volume, and temperature is in degrees Centigrade. It should be understood that variations in the methods as described may be made without changing the fundamental aspects that the Examples are meant to illustrate.

[0904] *Example 1*

[0905] The objective was to show that a Bayesian maximum likelihood estimation (MLE) algorithm that uses parent genotypes to calculate fetal fraction improves accuracy of non-invasive prenatal trisomy diagnosis compared to published methods.

[0906] Simulated sequencing data for maternal cfDNA was created by sampling reads obtained on trisomy-21 and respective mother cell lines. The rate of correct disomy and trisomy calls were determined from 500 simulations at various fetal fractions for a published method (Chiu et al. BMJ 2011;342:c7401) and our MLE-based algorithm. We validated the simulations by obtaining 5 million shotgun reads from four pregnant mothers and respective fathers collected under an IRB-approved protocol. Parental genotypes were obtained on a 290K SNP array. (See FIG. 14)

[0907] In simulations, the MLE-based approach achieved 99.0% accuracy for fetal fractions as low as 9% and reported confidences that corresponded well to overall accuracy. We validated these results using four real samples wherein we obtained all correct calls with a computed confidence exceeding 99%. In contrast, our implementation of the published algorithm for Chiu et al. required 18% fetal fraction to achieve 99.0% accuracy, and achieved only 87.8% accuracy at 9% fetal DNA.

[0908] Fetal fraction determination from parental genotypes in conjunction with a MLE-based approach achieves greater accuracy than published algorithms at the fetal fractions expected during the 1st and early 2nd trimester. Furthermore, the method disclosed herein produces a confidence metric that is crucial in determining the reliability of the result, especially at low fetal fractions where ploidy detection is more difficult. Published methods use a less accurate

threshold method for calling ploidy based on large sets of disomy training data, an approach that predefines a false positive rate. In addition, without a confidence metric, published methods are at risk of reporting false negative results when there is insufficient fetal cfDNA to make a call. In some embodiments, a confidence estimate is calculated for the called ploidy state.

[0909] *Example 2*

[0910] The objective was to improve non-invasive detection of fetal trisomy 18, 21, and X particularly in samples consisting of low fetal fraction by using a targeted sequencing approach combined with parent genotypes and Hapmap data in a Bayesian Maximum Likelihood Estimation (MLE) algorithm.

[0911] Maternal samples from four euploid and two trisomy-positive pregnancies and respective paternal samples were obtained under an IRB-approved protocol from patients where fetal karyotype was known. Maternal cfDNA was extracted from plasma and roughly 10 million sequence reads were obtained following preferential enrichment that targeted specific SNPs. Parent samples were similarly sequenced to obtain genotypes.

[0912] The described algorithm correctly called chromosome 18 and 21 disomy for all euploid samples and normal chromosomes of aneuploid samples. Trisomy 18 and 21 calls were correct, as were chromosome X copy numbers in male and female fetuses. The confidence produced by the algorithm was in excess of 98% in all cases.

[0913] The method described accurately reported the ploidy of all tested chromosomes from six samples, including samples comprised of less than 12% fetal DNA, which account for roughly 30% of 1st and early 2nd-trimester samples. The crucial difference between the instant MLE algorithm and published methods is that it leverages parent genotypes and Hapmap data to improve accuracy and generate a confidence metric. At low fetal fractions, all methods become less accurate; it is important to correctly identify samples without sufficient fetal cfDNA to make a reliable call. Others have used chromosome Y specific probes to estimate fetal fraction of male fetuses, but concurrent parental genotyping enables estimation of fetal fraction for both sexes. Another inherent limitation of published methods using untargeted shotgun sequencing is that accuracy of ploidy calling varies among chromosomes due to differences in factors such as GC richness. The instant targeted sequencing approach is largely independent of such chromosome-scale variations and yields more consistent performance between chromosomes.

[0914] *Example 3*

[0915] The objective was to determine if trisomy is detectable with high confidence on a triploid fetus, using novel informatics to analyze SNP loci of free floating fetal DNA in maternal plasma.

[0916] 20mL of blood was drawn from a pregnant patient following abnormal ultrasound. After centrifugation, maternal DNA was extracted from the buffy coat (DNEASY, QIAGEN); cell-free DNA was extracted from plasma (QIAAMP QIAGEN). Targeted sequencing was applied to SNP loci on chromosomes 2, 21, and X in both DNA samples. Maximum-Likelihood Bayesian estimation selected the most likely hypothesis from the set of all possible ploidy states. The method determines fetal DNA fraction, ploidy state and explicit confidences in the ploidy determination. No assumptions are made about the ploidy of a reference chromosome. The diagnostic uses a test statistic that is independent of sequence read counts, which is the recent state of the art.

[0917] The instant method accurately diagnosed trisomy of chromosomes 2 and 21. Child fraction was estimated at 11.9% [CI 11.7-12.1]. The fetus was found to have one maternal and two paternal copies of chromosomes 2 and 21 with confidence of effectively 1 (error probability $<10^{-30}$). This was achieved with 92,600 and 258,100 reads on chromosomes 2 and 21 respectively.

[0918] This is the first demonstration of non-invasive prenatal diagnosis of trisomic chromosomes from maternal blood where the fetus was triploid, as confirmed by metaphase karyotype. Extant methods of non-invasive diagnosis would not detect aneuploidy in this sample. Current methods rely on a surplus of sequence reads on a trisomic chromosome relative to disomic reference chromosomes; but a triploid fetus has no disomic reference. Furthermore, extant methods would not achieve similarly high-confidence ploidy determination with this fraction of fetal DNA and number of sequence reads. It is straightforward to extend the approach to all 24 chromosomes.

[0919] *Example 4*

[0920] The following protocol was used for 800-plex amplification of DNA isolated from maternal plasma from a euploid pregnancy and also genomic DNA from a triploidy 21 cell line using standard PCR (meaning no nesting was used). Library preparation and amplification

involved single tube blunt ending followed by A-tailing. Adaptor ligation was run using the ligation kit found in the AGILENT SURESELECT kit, and PCR was run for 7 cycles. Then, 15 cycles of STA (95°C for 30s; 72°C for 1 min; 60°C for 4 min; 65°C for 1 min; 72°C for 30s) using 800 different primer pairs targeting SNPs on chromosomes 2, 21 and X. The reaction was run with 12.5 nM primer concentration. The DNA was then sequenced with an ILLUMINA IIGAX sequencer. The sequencer output 1.9 million reads, of which 92% mapped to the genome; of those reads that mapped to the genome, more than 99% mapped to one of the regions targeted by the targeted primers. The numbers were essentially the same for both the plasma DNA and the genomic DNA. FIG. 15 shows the ratio of the two alleles for the ~780 SNPs that were detected by the sequencer in the genomic DNA that was taken from a cell line with known trisomy at chromosome 21. Note that the allele ratios are plotted here for ease of visualization, because the allele distributions are not straightforward to read visually. The circles represent SNPs on disomic chromosomes, while the stars represent SNPs on a trisomic chromosome. FIG. 16 is another representation of the same data as in FIG. 15, where the Y-axis is the relative number of A and B measured for each SNP, and where the X-axis is the SNP number where the SNPs are separated by chromosome. In FIG. 16, SNP 1 to 312 are found on chromosome 2, from SNP 313 to 605 are found on chromosome 21 which is trisomic, and from SNP 606 to 800 are on chromosome X. The data from chromosomes 2 and X show a disomic chromosome, as the relative sequence counts lie in three clusters: AA at the top of the graph, BB at the bottom of the graph, and AB in the middle of the graph. The data from chromosome 21, which is trisomic, shows four clusters: AAA at the top of the graph, AAB around the 0.65 line (2/3), ABB around the .35 line (1/3), and BBB at the bottom of the graph.

[0921] FIGs. 17A-D show data for the same 800-plex protocol, but measured on DNA that was amplified from four plasma samples from pregnant women. For these four samples, we expect to see seven clusters of dots: (1) along the top of the graph are those loci where both the mother and the fetus are AA, (2) slightly below the top of the graph are those loci where the mother is AA and the fetus is AB, (3) slightly above the 0.5 line are those loci where the mother is AB and the fetus is AA, (4) along the 0.5 line are those loci where the mother and the fetus are both AB, (5) slightly below the 0.5 line are those loci where the mother is AB and the fetus is BB, (6) slightly above the bottom of the graph are those loci where the mother is BB and the fetus is AB, (7) along the bottom of the graph are those loci where both the mother and the fetus are BB. The smaller the fetal fraction, the less the separation between clusters (1) and (2), between clusters

(3), (4) and (5), and between clusters (6) and (7). The separation is expected to be half of the fraction of DNA that is of fetal origin. For example if the DNA is 20% fetal, and 80% maternal, we expect (1) through (7) to be centered at 1.0, 0.9, 0.6, 0.5, 0.4, 0.1 and 0.0 respectively; see for example FIG. 17D, POOL1_BC5_ref_rate. If, instead the DNA is 8% fetal, and 92% maternal, we expect (1) through (7) to be centered at 1.00, 0.96, 0.54, 0.50, 0.46, 0.04 and 0.00 respectively; see for example FIG. 17B, POOL1_BC2_ref_rate. If there is not fetal DNA detected, we do not expect to see (2), (3), (5), or (6); alternately we could say that the separation is zero, and therefore (1) and (2) are on top of each other, as are (3), (4) and (5), and also (6) and (7); see *e.g.* FIG. 17C, POOL1_BC7_ref_rate. Note that the fetal fraction for FIG. 17A, POOL1_BC1_ref_rate is about 25%.

[0922] *Example 5*

[0923] Most methods of DNA amplification and measurement will produce some allele bias, wherein the two alleles that are typically found at a locus are detected with intensities or counts that are not representative of the actual amounts of alleles in the sample of DNA. For example, for a single individual, at a heterozygous locus we expect to see a 1:1 ratio of the two alleles, which is the theoretical ratio expected for a heterozygous locus; however due to allele bias, we may see 55:45, or even 60:40. Also note that in the context of sequencing, if the depth of read is low, then simple stochastic noise could result in significant allele bias. In an embodiment, it is possible to model the behavior of each SNP such that if a consistent bias is observed for particular alleles, this bias can be corrected for. FIG. 18 shows the fraction of data that can be explained by binomial variance, before and after bias correction. In FIG. 18, the stars represent the observed allele bias on raw sequence data for the 800-plex experiment; the circles represent the allele bias after correction. Note that if there were no allele bias at all, we would expect the data to fall along the $x=y$ line. A similar set of data that was produced by amplifying DNA using a 150-plex targeted amplification produced data that fell very closely on the 1:1 line after bias correction.

[0924] *Example 6*

[0925] Universal amplification of DNA using ligated adaptors with primers specific to the adaptor tags, where the primer annealing and extension times are limited to a few minutes has

the effect of enriching the proportion of shorter DNA strands. Most library protocols designed for creating DNA libraries suitable for sequencing contain such a step, and example protocols are published and well known to those in the art. In some embodiments of the invention, adaptors with a universal tag are ligated to the plasma DNA, and amplified using primers specific to the adaptor tag. In some embodiments, the universal tag can be the same tag as used for sequencing, it can be a universal tag only for PCR amplification, or it can be a set of tags. Since the fetal DNA is typically short in nature, while the maternal DNA can be both short and long in nature, this method has the effect of enriching the proportion of fetal DNA in the mixture. The free floating DNA, thought to be DNA from apoptotic cells, and which contains both fetal and maternal DNA, is short – mostly under 200 bp. Cellular DNA released by cell lysis, a common phenomenon after phlebotomy, is typically almost exclusively maternal, and is also quite long – mostly above 500 bp. Therefore, blood samples that have sat around for more than a few minutes will contain a mixture of short (fetal + maternal) and longer (maternal) DNA. Performing a universal amplification with relatively short extension times on maternal plasma followed by targeted amplification will tend to increase the relative proportion of fetal DNA when compared to the plasma that has been amplified using targeted amplification alone. This can be seen in FIG. 19 which shows the measured fetal percent when the input is plasma DNA (vertical axis) vs. the measured fetal percent when the input DNA is plasma DNA that has had a library prepared using the ILLUMINA GAIIX library preparation protocol. All the dots fall below the line, indicating that the library preparation step enriches the fraction of DNA that is of fetal origin. Two samples of plasma that were red, indicating hemolysis and therefore that there would be an increased amount of long maternal DNA present from cell lysis, show a particularly significant enrichment of fetal fraction when the library preparation is performed prior to targeted amplification. The method disclosed herein is particularly useful in cases where there is hemolysis or some other situation has occurred where cells comprising relatively long strands of contaminating DNA have lysed, contaminating the mixed sample of short DNA with the long DNA. Typically the relatively short annealing and extension times are between 30 seconds and 2 minutes, though they could be as short as 5 or 10 seconds or less, or as long as 5 or 10 minutes.

[0926] *Example 7*

[0927] The following protocol was used for 1,200-plex amplification of DNA isolated from maternal plasma from a euploid pregnancy and also genomic DNA from a triploidy 21 cell line using a direct PCR protocol, and also a semi-nested approach. Library preparation and amplification involved single tube blunt ending followed by A-tailing. Adaptor ligation was run using a modification of the ligation kit found in the AGILENT SURESELECT kit, and PCR was run for 7 cycles. In the targeted primer pool, there were 550 assays for SNPs from chromosome 21, and 325 assays for SNPs from each of chromosomes 1 and X. Both protocols involved 15 cycles of STA (95°C for 30s; 72°C for 1 min; 60°C for 4 min; 65°C for 30s; 72°C for 30s) using 16 nM primer concentration. The semi-nested PCR protocol involved a second amplification of 15 cycles of STA (95°C for 30s; 72°C for 1 min; 60°C for 4 min; 65°C for 30s; 72°C for 30s) using an inner forward tag concentration of 29 nM, and a reverse tag concentration of 1 uM or 0.1 uM. The DNA was then sequenced with an ILLUMINA HIGAX sequencer. For the direct PCR protocol, 73% of the reads map to the genome; for the semi-nested protocol, 97.2% of the sequence reads map to the genome. Therefore, the semi-nested protocol result in approximately 30% more information, presumably mostly due to the elimination of primers that are most likely to cause primer dimers.

[0928] The depth of read variability tends to be higher when using the semi-nested protocol than when the direct PCR protocol is used (see FIG. 20) where the diamonds refer to the depth of read for loci run with the semi-nested protocol, and the squares refer to the depth of read for loci run with no nesting. The SNPs are arranged by depth of read for the diamonds, so the diamonds all fall on a curved line, while the squares appear to be loosely correlated; the arrangements of the SNPs is arbitrary, and it is the height of the dot that denotes depth of read rather than its location left to right.

[0929] In some embodiments, the methods described herein can achieve excellent depth of read (DOR) variances. For example, in one version of this Example (FIG. 21) using a 1,200-plex direct PCR amplification of genomic DNA, of the 1,200 assays: 1186 assays had a DOR greater than 10; the average depth of read was 400; 1063 assays (88.6%) had a depth of read of between 200 and 800, and ideal window where the number of reads for each allele is high enough to give meaningful data, while the number of reads for each allele is not so high that the marginal use of those reads was particularly small. Only 12 alleles had higher depth of read with the highest at 1035 reads. The standard deviation of the DOR was 290, the average DOR was 453, the coefficient of variance of the DOR was 64%, there were 950,000 total reads, and 63.1% of the

reads mapped to the genome. In another experiment (FIG. 22) using a 1,200-plex semi-nested protocol, the DOR was higher. The standard deviation of the DOR was 583, the average DOR was 630, the coefficient of variance of the DOR was 93%, there were 870,000 total reads, and 96.3% of the reads mapped to the genome. Note, in both these cases, the SNPs are arranged by the depth of read for the mother, so the curved line represents the maternal depth of read. The differentiation between child and father is not significant; it is only the trend that is significant for the purpose of this explanation.

[0930] *Example 8*

[0931] In an experiment, the semi-nested 1,200-plex PCR protocol was used to amplify DNA from one cell and from three cells. This experiment is relevant to prenatal aneuploidy testing using fetal cells isolated from maternal blood, or for preimplantation genetic diagnosis using biopsied blastomeres or trophoctoderm samples. There were 3 replicates of 1 and 3 cells from 2 individuals (46 XY and 47 XX+21) per condition. Assays targeted chromosomes 1, 21 and X. Three different lysis methods were used: ARCTURUS, MPERv2 and Alkaline lysis. Sequencing was run multiplexing 48 samples in one sequencing lane. The algorithm returned correct ploidy calls for each of the three chromosomes, and for each of the replicates.

[0932] *Example 9*

[0933] In one experiment, four maternal plasma samples were prepared and amplified using a hemi-nested 9,600-plex protocol. The samples were prepared in the following way: Up to 40 mL of maternal blood were centrifuged to isolate the buffy coat and the plasma. The genomic DNA in the maternal sample was prepared from the buffy coat and paternal DNA was prepared from a blood sample or saliva sample. Cell-free DNA in the maternal plasma was isolated using the QIAGEN CIRCULATING NUCLEIC ACID kit and eluted in 45 uL TE buffer according to manufacturer's instructions. Universal ligation adapters were appended to the end of each molecule of 35 uL of purified plasma DNA and libraries were amplified for 7 cycles using adaptor specific primers. Libraries were purified with AGENCOURT AMPURE beads and eluted in 50 ul water.

[0934] 3 ul of the DNA was amplified with 15 cycles of STA (95°C for 10 min for initial polymerase activation, then 15 cycles of 95°C for 30s; 72°C for 10 s; 65°C for 1 min; 60°C for

8 min; 65°C for 3 min and 72°C for 30s; and a final extension at 72°C for 2 min) using 14.5 nM primer concentration of 9600 target-specific tagged reverse primers and one library adaptor specific forward primer at 500 nM.

[0935] The hemi-nested PCR protocol involved a second amplification of a dilution of the first STAs product for 15 cycles of STA (95°C for 10 min for initial polymerase activation, then 15 cycles of 95°C for 30s; 65°C for 1 min; 60°C for 5 min; 65°C for 5 min and 72°C for 30s; and a final extension at 72°C for 2 min) using reverse tag concentration of 1000 nM, and a concentration of 16.6 u nM for each of 9600 target-specific forward primers.

[0936] An aliquot of the STA products was then amplified by standard PCR for 10 cycles with 1 uM of tag-specific forward and barcoded reverse primers to generate barcoded sequencing libraries. An aliquot of each library was mixed with libraries of different barcodes and purified using a spin column.

[0937] In this way, 9,600 primers were used in the single-well reactions; the primers were designed to target SNPs found on chromosomes 1, 2, 13, 18, 21, X and Y. The amplicons were then sequenced using an ILLUMINA GAIIIX sequencer. Per sample, approximately 3.9 million reads were generated by the sequencer, with 3.7 million reads mapping to the genome (94%), and of those, 2.9 million reads (74%) mapped to targeted SNPs with an average depth of read of 344 and a median depth of read of 255. The fetal fraction for the four samples was found to be 9.9%, 18.9%, 16.3%, and 21.2%

[0938] Relevant maternal and paternal genomic DNA samples amplified using a semi-nested 9600-plex protocol and sequenced. The semi-nested protocol is different in that it applies 9,600 outer forward primers and tagged reverse primers at 7.3 nM in the first STA. Thermocycling conditions and composition of the second STA, and the barcoding PCR were the same as for the hemi-nested protocol.

[0939] The sequencing data was analyzed using informatics methods disclosed herein and the ploidy state was called at six chromosomes for the fetuses whose DNA was present in the 4 maternal plasma samples. The ploidy calls for all 28 chromosomes in the set were called correctly with confidences above 99.2% except for one chromosome that was called correctly, but with a confidence of 83%.

[0940] FIG. 23 shows the depth of read of the 9,600-plex hemi-nesting approach along with the depth of read of the 1,200-plex semi-nested approach described in Example 7, though the number of SNPs with a depth of read greater than 100, greater than 200 and greater than 400

was significantly higher than in the 1,200-plex protocol. The number of reads at the 90th percentile can be divided by the number of reads at the 10th percentile to give a dimensionless metric that is indicative of the uniformity of the depth of read; the smaller the number, the more uniform (narrow) the depth of read. The average 90th percentile/10th percentile ratio is 11.5 for the method run in Example 9, while it is 5.6 for the method run in Example 7. A narrower depth of read for a given protocol plexity is better for sequencing efficiency, as fewer sequence reads are necessary to ensure that a certain percentage of reads are above a read number threshold.

[0941] *Example 10*

[0942] In one experiment, four maternal plasma samples were prepared and amplified using a semi-nested 9,600-plex protocol. Details of Example 10 were very similar to Example 9, the exception being the nesting protocol, and including the identity of the four samples. The ploidy calls for all 28 chromosomes in the set were called correctly with confidences above 99.7%. 7.6 million (97%) of reads mapped to the genome, and 6.3 million (80%) of the reads mapped to the targeted SNPs. The average depth of read was 751, and the median depth of read was 396.

[0943] *Example 11*

[0944] In one experiment, three maternal plasma samples were split into five equal portions, and each portion was amplified using either 2,400 multiplexed primers (four portions) or 1,200 multiplexed primers (one portion) and amplified using a semi-nested protocol, for a total of 10,800 primers. After amplification, the portions were pooled together for sequencing. Details of Example 11 were very similar to Example 9, the exception being the nesting protocol, and the split and pool approach. The ploidy calls for all 21 chromosomes in the set were called correctly with confidences above 99.7%, except for one missed call where the confidence was 83%. 3.4 million reads mapped to targeted SNPs, the average depth of read was 404 and the median depth of read was 258.

[0945] *Example 12*

[0946] In one experiment, four maternal plasma samples were split into four equal portions, and each portion was amplified using 2,400 multiplexed primers and amplified using a semi-nested

protocol, for a total of 9,600 primers. After amplification, the portions were pooled together for sequencing. Details of Example 12 were very similar to Example 9, the exception being the nesting protocol, and the split and pool approach. The ploidy calls for all 28 chromosomes in the set were called correctly with confidences above 97%, except for one missed call where the confidence was 78%. 4.5 million reads mapped to targeted SNPs, the average depth of read was 535 and the median depth of read was 412.

[0947] *Example 13*

[0948] In one experiment, four maternal plasma samples were prepared and amplified using a 9,600-plex triply hemi-nested protocol, for a total of 9,600 primers. Details of Example 12 were very similar to Example 9, the exception being the nesting protocol which involved three rounds of amplification; the three rounds involved 15, 10 and 15 STA cycles respectively. The ploidy calls for 27 of 28 chromosomes in the set were called correctly with confidences above 99.9%, except for one that was called correctly with 94.6%, and one missed call with a confidence of 80.8%. 3.5 million reads mapped to targeted SNPs, the average depth of read was 414 and the median depth of read was 249.

[0949]

[0950] *Example 14*

[0951] In one Example 45 sets of cells were amplified using a 1,200-plex semi-nested protocol, sequenced, and ploidy determinations were made at three chromosomes. Note that this experiment is meant to simulate the conditions of performing pre-implantation genetic diagnosis on single-cell biopsies from day 3 embryos, or trophoctoderm biopsies from day 5 embryos. 15 individual single cells and 30 sets of three cells were placed in 45 individual reaction tubes for a total of 45 reactions where each reaction contained cells from only one cell line, but the different reactions contained cells from different cell lines. The cells were prepared into 5 ul washing buffer and lysed the by adding 5 ul ARCTURUS PICOPURE lysis buffer (APPLIED BIOSYSTEMS) and incubating at 56°C for 20 min, 95°C for 10 min..

[0952] The DNA of the single/three cells was amplified with 25 cycles of STA (95°C for 10 min for initial polymerase activation, then 25 cycles of 95°C for 30s; 72°C for 10 s; 65°C for 1 min; 60°C for 8 min; 65°C for 3 min and 72°C for 30s; and a final extension at 72°C for 2 min) using 50 nM primer concentration of 1200 target-specific forward and tagged reverse primers.

[0953] The semi-nested PCR protocol involved three parallel second amplification of a dilution of the first STAs product for 20 cycles of STA (95°C for 10 min for initial polymerase activation, then 15 cycles of 95°C for 30s; 65°C for 1 min; 60°C for 5 min; 65°C for 5 min and 72°C for 30s; and a final extension at 72°C for 2 min) using reverse tag specific primer concentration of 1000 nM, and a concentration of 60 nM for each of 400 target-specific nested forward primers. In the three parallel 400-plex reactions the total of 1200 targets amplified in the first STA were thus amplified.

[0954] An aliquot of the STA products was then amplified by standard PCR for 15 cycles with 1 uM of tag-specific forward and barcoded reverse primers to generate barcoded sequencing libraries. An aliquot of each library was mixed with libraries of different barcodes and purified using a spin column.

[0955] In this way, 1,200 primers were used in the single cell reactions; the primers were designed to target SNPs found on chromosomes 1, 21 and X. The amplicons were then sequenced using an ILLUMINA GAIIX sequencer. Per sample, approximately 3.9 million reads were generated by the sequencer, with 500,000 to 800,000 million reads mapping to the genome (74% to 94% of all reads per sample).

[0956] Relevant maternal and paternal genomic DNA samples from cell lines were analyzed using the same semi-nested 1200-plex assay pool with a similar protocol with fewer cycles and 1200-plex second STA, and sequenced.

[0957] The sequencing data was analyzed using informatics methods disclosed herein and the ploidy state was called at the three chromosomes for the samples.

[0958] FIG. 24 shows normalized depth of read ratios (vertical axis) for six samples at three chromosomes (1 = chrom 1; 2 = chrom 21; 3 = chrom X). The ratios were set to be equal to the number of reads mapping to that chromosome, normalized, and divided by the number of reads mapping to that chromosome averaged over three wells each comprising three 46XY cells. The three sets of data points corresponding to the 46XY reactions are expected to have ratios of 1:1. The three sets of data points corresponding to the 47XX+21 cells are expected to have ratios of 1:1 for chromosome 1, 1.5:1 for chromosome 21, and 2:1 for chromosome X.

[0959] FIG. 25A-25C show allele ratios plotted for three chromosomes (1, 21, X) for three reaction. The reaction in the lower left shows a reaction on three 46XY cells (FIG. 25B). The left region are the allele ratios for chromosome 1, the middle region are the allele ratios for chromosome 21, and the right region are the allele ratios for chromosome X. For the 46XY

cells, for chromosome 1 we expect to see ratios of 1, 0.5 and 0, corresponding to AA, AB and BB SNP genotypes. For the 46XY cells, for chromosome 21 we expect to see ratios of 1, 0.5 and 0, corresponding to AA, AB and BB SNP genotypes. For the 46XY cells, for chromosome X we expect to see ratios of 1 and 0, corresponding to A, and B SNP genotypes. The reaction in the lower right shows a reaction on three 47XX+21 cells (FIG. 25C). The allele ratios are segregated by chromosome as in the lower left graph. For the 47XX+21 cells, for chromosome 1 we expect to see ratios of 1, 0.5 and 0, corresponding to AA, AB and BB SNP genotypes. For the 47XX+21 cells, for chromosome 21 we expect to see ratios of 1, 0.67, 0.33 and 0, corresponding to AAA, AAB, ABB and BBB SNP genotypes. For the 47XX+21 cells, for chromosome X we expect to see ratios of 1, 0.5 and 0, corresponding to AA, AB, and BB SNP genotypes. The plot in the upper right was made on a reaction comprising 1 ng of genomic DNA from the 47XX+21 cell line (FIG. 25A). FIGs. 25B and 26B shows the same graphs as in FIG. 25A-25C, but for reactions performed on only one cell. The left graph was a reaction that contained a 47XX+21 cell (FIG. 26A), and the right graph was for a reaction that contained a 46XX cell (FIG. 26B).

[0960] From the graphs shown in FIGs. 25A-25C and FIGs. 26A and 26B, it is visually apparent that there are two clusters of dots for chromosomes where we expect to see ratios of 1 and 0; three clusters of dots for chromosomes where we expect to see ratios of 1, 0.5, and 0, and four clusters of dots for chromosomes where we expect to see ratios of 1, 0.67, 0.33 and 0. The parental support algorithm was able to make correct calls on all of the three chromosomes for all of the 45 reactions.

[0961] *Example 15*

[0962] In one experiment, maternal plasma samples were prepared and amplified using a hemi-nested 19,488-plex protocol. The samples were prepared in the following way: up to 20 mL of maternal blood were centrifuged to isolate the buffy coat and the plasma. The genomic DNA in the maternal sample was prepared from the buffy coat and paternal DNA was prepared from a blood sample or saliva sample. Cell-free DNA in the maternal plasma was isolated using the QIAGEN CIRCULATING NUCLEIC ACID kit and eluted in 50 uL TE buffer according to manufacturer's instructions. Universal ligation adapters were appended to the end of each molecule of 40 uL of purified plasma DNA and libraries were amplified for 9 cycles using

adaptor specific primers. Libraries were purified with AGENCOURT AMPURE beads and eluted in 50 ul DNA suspension buffer.

[0963] 6 ul of the DNA was amplified with 15 cycles of STAR 1 (95°C for 10 min for initial polymerase activation, then 15 cycles of 96°C for 30s; 65°C for 1 min; 58°C for 6 min; 60°C for 8 min; 65°C for 4 min and 72°C for 30s; and a final extension at 72°C for 2 min) using 7.5 nM primer concentration of 19,488 target-specific tagged reverse primers and one library adaptor specific forward primer at 500 nM.

[0964] The hemi-nested PCR protocol involved a second amplification of a dilution of the STAR 1 product for 15 cycles (STAR 2) (95°C for 10 min for initial polymerase activation, then 15 cycles of 95°C for 30s; 65°C for 1 min; 60°C for 5 min; 65°C for 5 min and 72°C for 30s; and a final extension at 72°C for 2 min) using reverse tag concentration of 1000 nM, and a concentration of 20 nM for each of 19,488 target-specific forward primers.

[0965] An aliquot of the STAR 2 products was then amplified by standard PCR for 12 cycles with 1 uM of tag-specific forward and barcoded reverse primers to generate barcoded sequencing libraries. An aliquot of each library was mixed with libraries of different barcodes and purified using a spin column.

[0966] In this way, 19,488 primers were used in the single-well reactions; the primers were designed to target SNPs found on chromosomes 1, 2, 13, 18, 21, X and Y. The amplicons were then sequenced using an ILLUMINA GAIIX sequencer. For plasma samples, approximately 10 million reads were generated by the sequencer, with 9.4-9.6 million reads mapping to the genome (94-96 %), and of those, 99.95 % mapped to targeted SNPs with a mean depth of read of 460 and a median depth of read of 350. For comparison, a perfectly even distribution would be: 10M reads / 19,488 targets = 513 reads/target. For primer-dimers, 30,000 reads were from sequenced primer-dimers (0.3% of the reads generated by the sequencer). For genomic samples, 99.4-99.7% of the reads mapped to the genome, of those, 99.99% of the mapped to targeted SNPs, and 0.1% of the reads generated by the sequencer were primer-dimers.

[0967] For plasma samples with 10 million sequencing reads, typically at least 19,350 of the 19,488 targeted SNPs (99.3 %) are amplified and sequenced. For DNA samples with 2M sequencing reads, typically at least 19,000 targeted SNPs (97.5%) are amplified and sequenced. The lower number may be due to sampling noise since the number of reads is lower and the sequencer misses some of the amplified products. If desired, the number of sequencing reads can be increased to increase the number of targeted SNPs that are amplified and sequenced.

[0968] Relevant maternal and paternal genomic DNA samples amplified using a semi-nested 19,488 outer forward primers and tagged reverse primers at 7.5 nM in the STAR 1. Thermocycling conditions and composition of STAR 2, and the barcoding PCR were the same as for the hemi-nested protocol.

[0969] The average fetal fraction for 407 samples was found to be 14.8%. The sequencing data was analyzed using informatics methods disclosed herein and the ploidy state was called at four chromosomes (13, 18, 21, Y) for the fetuses whose DNA was present in 378 of the 407 maternal plasma samples, and at chromosome X in 375 of the 407 maternal plasma samples. The ploidy calls for all 1,887 chromosomes in the set were called correctly with confidences above 90%. 1882 of the 1887 calls were above 95%; and 1,862 of the 1,887 calls were called with confidences above 99%.

[0970] A similar control experiment was performed using water instead of DNA extracted from plasma in the plasma PCR protocol. Based on six such trials of an experiment, 5-6% of the sequenced reads were primer-dimers. Other sequenced reads were due to background noise. This experiment demonstrates that even in the absence of a nucleic acid sample with target loci for the primers to hybridize to (rather than hybridizing to other primers and forming amplified primer dimers) few primer dimers are formed.

[0971] *Example 16*

[0972] The following Example illustrates an exemplary method for designing and selecting a library of primers that can be used in any of the multiplexed PCR methods of the invention. The goal is to select primers from an initial library of candidate primers that can be used to simultaneously amplify a large number of target loci (or a subset of target loci) in a single reaction. For an initial set of candidate target loci, primers did not have to be designed or selected for each target locus. Preferably, primers are designed and selected for a large portion of the most desirable target loci.

[0973] *Step 1*

[0974] A set of candidate target loci (such as SNPs) were selected based on publically available information about desired parameters for the target loci, such as frequency of the SNPs within a target population or heterozygosity rate of the SNPs (worldwide web at

ncbi.nlm.nih.gov/projects/SNP/; Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001 Jan 1;29(1):308-11, which are each incorporated by reference in its entirety). For each candidate locus, one or more PCR primer pairs were designed using the Primer3 program (the worldwide web at primer3.sourceforge.net; libprimer3 release 2.2.3, which is hereby incorporated by reference in its entirety). If there were no feasible designs for PCR primers for a particular target locus, then that target locus was eliminated from further consideration. If desired, a "target locus score" (higher score representing higher desirability) can be calculated for most or all of the target loci, such as a target locus score calculated based on a weighted average of various desired parameters for the target loci. The parameters may be assigned different weights based on their importance for the particular application that the primers will be used for. Exemplary parameters include the heterozygosity rate of the target locus, the disease prevalence associated with a sequence (*e.g.*, a polymorphism) at the target locus, the disease penetrance associated with a sequence (*e.g.*, a polymorphism) at the target locus, the specificity of the candidate primer(s) used to amplify the target locus, the size of the candidate primer(s) used to amplify the target locus, and the size of the target amplicon. In some embodiments, the specificity of the candidate primer for the target locus includes the likelihood that the candidate primer will mis-prime by binding and amplifying a locus other than the target locus it was designed to amplify. In some embodiments, one or more or all the candidate primers that mis-prime are removed from the library.

[0975] *Step 2*

[0976] A thermodynamic interaction score was calculated between each primer and all primers for all other target loci from Step 1 (*see, e.g.*, Allawi, H. T. & SantaLucia, J., Jr. (1998), "Thermodynamics of Internal C-T Mismatches in DNA", *Nucleic Acids Res.* 26, 2694-2701; Peyret, N., Seneviratne, P. A., Allawi, H. T. & SantaLucia, J., Jr. (1999), "Nearest-Neighbor Thermodynamics and NMR of DNA Sequences with Internal A-A, C-C, G-G, and T-T Mismatches", *Biochemistry* 38, 3468-3477; Allawi, H. T. & SantaLucia, J., Jr. (1998), "Nearest-Neighbor Thermodynamics of Internal A-C Mismatches in DNA: Sequence Dependence and pH Effects", *Biochemistry* 37, 9435-9444.; Allawi, H. T. & SantaLucia, J., Jr. (1998), "Nearest Neighbor Thermodynamic Parameters for Internal G-A Mismatches in DNA", *Biochemistry* 37, 2170-2179; and Allawi, H. T. & SantaLucia, J., Jr. (1997), "Thermodynamics and NMR of

Internal G-T Mismatches in DNA", *Biochemistry* 36, 10581-10594; MultiPLX 2.1 (Kaplinski L, Andreson R, Puurand T, Remm M. MultiPLX: automatic grouping and evaluation of PCR primers. *Bioinformatics*. 2005 Apr 15;21(8):1701-2, which are each hereby incorporated by reference in its entirety). This step resulted in a 2D matrix of interaction scores. The interaction score predicted the likelihood of primer-dimers involving the two interacting primers. The score was calculated as follows:

[0977] $\text{interaction_score} = \max(-\text{deltaG_2}, 0.8 * (-\text{deltaG_1}))$

[0978] where

[0979] deltaG_2 = Gibbs energy (energy required to break the dimer) for a dimer that is extensible by PCR on both ends, *i.e.*, the 3' end of each primer anneals to the other primer; and

[0980] deltaG_1 = Gibbs energy for a dimer that is extensible by PCR on at least one end.

[0981] *Step 3:*

[0982] For each target locus, if there was more than one primer-pair design, then one design was selected using the following method:

[0983] For each primer-pair design for the locus, find the worst-case (highest) interaction score for the two primers in that design and all primers from all designs for all other target loci.

[0984] Pick the design with the best (lowest) worst-case interaction score.

[0985] *Step 4*

[0986] A graph was built such that each node represented one locus and its associated primer-pair design (*e.g.*, a Maximal Clique problem). One edge was created between every pair of nodes. A weight was assigned to each edge equal to the worst-case (highest) interaction score between the primers associated with the two nodes connected by the edge.

[0987] *Step 5*

[0988] If desired, for every pair of designs for two different target loci where one of the primers from one design and one of the primers from the other design would anneal to overlapping target

regions, an additional edge was added between the nodes for the two design. The weight of these edges was set equal to the highest weight assigned in Step 4. Thus, Step 5 prevents the library from having primers that would anneal to overlapping target regions, and thus interfere with each other during a multiplex PCR reaction.

[0989] *Step 6*

[0990] An initial interaction score threshold was calculated as follows:

[0991] $\text{weight_threshold} = \max(\text{edge_weight}) - 0.05 * (\max(\text{edge_weight}) - \min(\text{edge_weight}))$

[0992] where

[0993] $\max(\text{edge_weight})$ is the maximum edge weight in the graph; and

[0994] $\min(\text{edge_weight})$ is the minimum edge weight in the graph.

[0995] The initial bounds for the threshold were set as follows:

[0996] $\max_weight_threshold = \max(\text{edge_weight})$

[0997] $\min_weight_threshold = \min(\text{edge_weight})$

[0998] *Step 7*

[0999] A new graph was constructed consisting of the same set of nodes as the graph from Step 5, only including edges with weights that exceed weight_threshold . Thus, step ignores interactions with scores equal to or below weight_threshold .

[1000] *Step 8*

[1001] Nodes (and all of the edges connected to the removed nodes) were removed from the graph of Step 7 until there were no edges left. Nodes were removed by applying the following procedure repeatedly:

[1002] Find the node with the highest degree (highest number of edges). If there is more than one then pick one arbitrarily.

[1003] Define the set of nodes consisting of the node picked above and all of the nodes connected to it, but excluding any nodes that have degree less than the node picked above.

[1004] Choose the node from the set that has the lowest target locus score (lower score representing lower desirability) from Step 1. Remove that node from the graph.

[1005] *Step 9*

[1006] If the number of nodes remaining in the graph satisfies the required number of target loci for the multiplexed PCR pool (within an acceptable tolerance), then the method was continued at Step 10.

[1007] If there were too many or too few nodes remaining in the graph, then a binary search was performed to determine what threshold values would result in the desired number of nodes remaining in the graphs. If there were too many nodes in the graph then, the weight threshold bounds were adjusted as follows:

[1008] $\text{max_weight_threshold} = \text{weight_threshold}$

[1009] Otherwise (if there are too few nodes in the graph), then the weight threshold bounds were adjusted as follows:

[1010] $\text{min_weight_threshold} = \text{weight_threshold}$

[1011] Then, the weight threshold was adjusted follows:

[1012] $\text{weight_threshold} = (\text{max_weight_threshold} + \text{min_weight_threshold}) / 2$

[1013] Steps 7-9 were repeated.

[1014] *Step 10*

[1015] The primer-pair designs associated with the nodes remaining in the graph were selected for the library of primers. This primer library can be used in any of the methods of the invention.

[1016] If desired, this method of designing and selecting primers can be performed for primer libraries in which only one primer (instead of a primer pair) is used for amplification of a target locus. In this case, a node presents one primer per target locus (rather than a primer pair).

[1017] *Example 17*

[1018] FIG. 27 is a graph comparing two primer libraries designed using the methods of the invention. This graph shows the number of loci with a particular minor allele frequency that are targeted by each primer library. During the selection of the “new pool” library, more primers were retained. This library enables the amplification of more target loci, especially target loci with relatively large minor allele frequencies (which are the more informative alleles for some method of the invention, such as for detecting fetal chromosomal abnormalities).

[1019] These primer libraries were used in the following multiplex PCR method. Blood (20-40 mL) was collected from each subject into two to four CELL-FREE™ DNA tubes (Streck). Plasma (a minimum of 7 mL) was isolated from each sample via a double centrifugation protocol of 2,000 g for 20 min, followed by 3,220 g for 30 min, with supernatant transfer following the first spin. cfDNA was isolated from 7-20 mL plasma using the QIAGEN QIAamp Circulating Nucleic Acid kit and eluted in 45 uL TE buffer. Pure maternal genomic DNA was isolated from the buffy coat obtained following the first centrifugation, and pure paternal genomic DNA was prepared similarly from a blood, saliva or buccal sample.

[1020] Maternal cfDNA, maternal genomic DNA, and paternal genomic DNA samples were pre-amplified for 15 cycles using 11,000 target-specific assays and an aliquot was transferred to a second PCR reaction of 15 cycles using nested primers. Finally, samples were prepared for sequencing by adding barcoded tags in a third 12-cycle round of PCR. Thus, 11,000 targets were amplified in a single reaction; the targets included SNPs found on chromosomes 13, 18, 21, X, and Y. The amplicons were then sequenced using an ILLUMINA GAIIx or HISEQ sequencer. Parental genotypes were sequenced at a lower read depth (~20% of cfDNA read depth) than the fetal genotypes.

[1021] *Example 18*

[1022] If desired, the size and quantity of the PCR products can be analyzed using standard methods, such as the use of the Agilent Technologies 2100 Bioanalyzer (FIG. 28A-M). For example, direct PCR methods described herein without nesting were used in 2,400-plex (FIGs. 28B-28G) and 19,488-plex experiments (FIGs. 28H to 28M). The amount of primer was 10 nM for FIGs. 28B-28D and 28H to 28J. The amount of primer was 1 nM for FIGs. 28E-28G and 28K to 28M. The amount of input DNA was 24 ng for FIGs. 28B, 28E, 28H, and 28K; 80 ng for FIGs. 28C, 28F, 28I, and 28L; and 250 ng for FIGs. 28D, 28G, 28J, and 28M. More input DNA resulted in a greater proportion of the desired 180 base pair product. The peak at 140 base pairs is a primer dimer product.

[1023] *Example 19*

[1024] A proof-of-principle study demonstrated the detection of T13, T18, T21, 45,X, and 47,XXY with equally high accuracies across all chromosomes.

[1025] *Patients*

[1026] Pregnant couples were enrolled at specific prenatal care centers under protocols approved by an Institutional Review Board pursuant to local laws. Inclusion criteria were at least 18 years of age, gestational age of at least nine weeks, singleton pregnancies, and signed informed consent. Blood samples were drawn from pregnant mothers, and a blood or buccal sample was collected from the father. Samples from 2 pregnancies with T13 (Patau Syndrome), 2 with T18 (Edwards Syndrome), 2 with T21 (Down's Syndrome), 2 with 45,X, 2 with 47,XXY, and 90 normal pregnancies were selected prior to testing from a cohort of ~500 women to test which chromosomal abnormalities the method detects. Normal fetal karyotype was confirmed by molecular karyotyping for the samples where post-birth child tissue was available. Euploid sample were drawn prior to invasive testing from low-risk women. Aneuploid samples were drawn at least 7 days after invasive testing and aneuploidy was confirmed via cytogenetic karyotyping or fluorescence *in situ* hybridization at independent laboratories.

[1027] *Sample Preparation and Multiplex PCR*

[1028] For the data in FIGs. 30A-E, 30G, 30H, and 31A-31G, sample preparation and 19,488-plex-PCR were performed as described in Example 15. For the data in FIG. 30F, sample preparation and 11,000-plex-PCR were performed as described in Example 17.

[1029]*Methodology and Data Analysis*

[1030] The algorithm considers parental genotypes and crossover frequency data (such as data from the HapMap database) to calculate expected allele distributions for 19,488 polymorphic loci for a very large number possible fetal ploidy states, and at various fetal cfDNA fractions. (FIGs. 29A-29C). Unlike allele ratio based-methods, it also takes into account linkage disequilibrium, and uses non-Gaussian data models to describe the expected distribution of allele measurements at a SNP given observed platform characteristics and amplification biases. It then compares the various predicted allele distributions to the actual allelic distributions as measured in the cfDNA sample (FIG. 29C), and calculates the likelihood of each hypothesis (monosomy, disomy, or trisomy, for which there are numerous hypotheses based on the various potential crossovers) based on the sequencing data. The algorithm sums the likelihoods of each individual

monosomy, disomy, or trisomy hypotheses (FIG. 29D), and calls the ploidy state with the maximum overall likelihood as the copy number and fetal fraction (FIG. 29E). Although laboratory researchers were not blinded to sample karyotype, the algorithm called the ploidy states without human intervention and was blind to the truth.

[1031] *Data Interpretation*

[1032] *Graphical Representations of the Generated Data*

[1033] To determine the ploidy state of a chromosome of interest, the algorithm considers the distribution of sequence counts from each of two possible alleles at 3,000 to 4,000 SNPs per chromosome. It is important to note that the algorithm makes ploidy calls using an approach that does not lend itself to visualization. Thus, for the purposes of illustration, the data is displayed here in a simplified fashion as ratios of the two most likely alleles, labeled as A and B, so that the relevant trends can be more readily visualized. This simplified illustration does not take into account some of the features of the algorithm. For example, two important aspects of the algorithm that are not possible to illustrate with a method of visualization that displays allele ratios are: 1) the ability to leverage linkage disequilibrium, *i.e.* the influence that a measurement at one SNP has on the likely identity of a neighboring SNP, and 2) the use of non-Gaussian data models that describe the expected distribution of allele measurements at a SNP given platform characteristics and amplification biases. Also note that the algorithm only considers the two most common alleles at each SNP, ignoring other possible alleles.

[1034] The graphical representations in FIG. 30A-30H include samples for which two, one, or three fetal chromosomes are present. Generally, these indicate euploidy (FIGs. 30A-30C) monosomy (FIG. 30D), and trisomy (FIGs. 30E-30H), respectively. In all plots, each spot represents a single SNP, where the targeted SNPs are plotted sequentially from left to right for one chromosome along the horizontal axes. The vertical axes indicate the number of reads for the A allele as a fraction of the total number of reads for both the A and B alleles for that SNP. Note that the measurements are made on total cfDNA isolated from maternal blood, and the cfDNA includes both maternal and fetal cfDNA; thus, each spot represents the combination of the fetal and maternal DNA contribution for that SNP. Therefore, increasing the proportion of maternal cfDNA from 0% to 100% will gradually shift some spots up or down within the plots, depending on the maternal and fetal genotype. This is described in more detail below with the corresponding plots.

[1035] If desired to facilitate visualization, the spots may be color-coded according to maternal genotype, as maternal genotype contributes more to the localization of each spot and the majority of trisomies are maternally-inherited; this assists in visualizing ploidy states. Specifically, SNPs for which the maternal genotype is AA may be indicated in red, those for which the maternal genotype is AB may be indicated in green, and those for which the maternal genotype is BB may be indicated in blue.

[1036] In all cases, SNPs that are homozygous for the A allele (AA) in both the mother and the fetus are found tightly associated with the upper limit of the plots, as the fraction of A allele reads is high because there should be no B alleles present. Conversely, SNPs that are homozygous for the B allele in both the mother and the fetus are found tightly associated with the lower limit of the plots, as the fraction of A allele reads is low because there should be only B alleles. The spots that are not tightly associated with the upper and lower limits of the plots represent SNPs for which the mother, the fetus, or both are heterozygous; these spots are useful for identifying fetal ploidy, but can also be informative for determining paternal versus maternal inheritance. These spots segregate based on both maternal and fetal genotypes and fetal fraction, and as such the precise position of each individual spot along the y-axis depends on both stoichiometry and fetal fraction. For example, loci where the mother is AA and the fetus is AB are expected to have a different fraction of A allele reads, and thus different positioning along the y-axis, depending on the fetal fraction.

[1037] *Two Chromosomes Present*

[1038] FIGs. 30A-30C depict data that indicate the presence of two chromosomes when the sample is entirely maternal (no fetal cfDNA present, FIG. 30A), contains a moderate fetal cfDNA fraction (FIG. 30B), or contains a high fetal cfDNA fraction (FIG. 30C).

[1039] FIG. 30A shows data obtained from cfDNA isolated from the blood of a non-pregnant woman. When there is no fetal cfDNA present and the sample contains only maternal cfDNA, the plots represent purely the euploid maternal genotype; the hallmark pattern includes “clusters” of spots: a red cluster tightly associated with the top of the plot (SNPs where the maternal genotype is AA), a blue cluster tightly associated with the bottom of the plot (SNPs where the maternal genotype is BB), and a single, centered green cluster (SNPs where the maternal genotype is AB).

[1040] When fetal cfDNA is present, the location of the spots shifts such that the clusters segregate into discrete “bands”. Note that for samples with a fetal fraction of 0%, the groupings of spots are referred to as “clusters” (as in FIG. 30A), and for all samples with a fetal fraction of >0%, the groupings of spots are referred to as “bands” (as in FIGs. 30B-30J). If the fetal fraction is high enough, these discrete bands will be readily visible. Specifically, FIGs. 30B and 30C demonstrate the characteristic pattern associated with two fetal chromosomes present at moderate and high fetal fractions, respectively. This pattern includes three central green bands that correspond to SNPs that are heterozygous in the mother, and two “peripheral” bands each at both the top (red) and bottom (blue) of the plots that correspond to SNPs that are homozygous in the mother.

[1041] FIG. 30B shows data obtained from cfDNA isolated from a plasma sample from a woman carrying a euploid fetus and with a 12% fetal cfDNA fraction. Here, the clusters of spots tightly associated with the top and bottom of the plot segregate into two discrete bands each: one red and one blue external peripheral band that remains tightly associated with the upper or lower limit of the plots, and one red and one blue internal peripheral band that has separated from the limits of the plots. These internal peripheral bands, centered around 0.92 and 0.08, represent SNPs for which the maternal genotype is AA and the fetal genotype is AB (indicated in red), and SNPs for which the maternal genotype is BB and the fetal genotype is AB (indicated in blue), respectively. The center cluster of green spots broadens, but at this fetal fraction the segregation into distinct bands is not readily visible.

[1042] At a high fetal cfDNA fraction, the typical pattern that indicates the presence of two chromosomes (a trio of green bands as well as two red and two blue peripheral bands) is readily apparent. FIG. 30C displays data obtained from a plasma sample from a woman carrying a euploid fetus at a fetal cfDNA fraction of 26%. Here, the peripheral bands have separated such that the internal band has shifted towards the center of the plot due to the altered levels of B alleles from the increased fetal cfDNA fraction. Significantly, at higher fetal fractions, the separation of the center green cluster into three distinct bands is now readily apparent. This central trio of bands, in this case clustering around 0.37, 0.50 and 0.63, corresponds to those SNPs where the maternal genotype is AB, and the fetal genotype is AA (top), AB (middle) and BB (bottom).

[1043] These hallmark patterns, namely three green bands and four peripheral bands (two red and two blue), indicate the presence of two chromosomes, as in autosomal euploidy or for the X chromosome in a female (XX) fetus.

[1044] *One Chromosome Present*

[1045] When the fetus only inherits a single chromosome, and thus only inherits a single allele, heterozygosity of the fetus is not possible. As such, the only possible fetal SNP identities are A or B. Thus, maternally-inherited monosomic chromosomes have a characteristic pattern of two central green bands that represent SNPs for which the mother is heterozygous, and only have single peripheral red and blue bands that represent SNPs for which the mother is homozygous, and which remain tightly associated with the upper and lower limits of the plots (1 and 0), respectively (FIG. 30D). Note the absence of internal peripheral bands. This pattern indicates the presence of one chromosome, as in maternally-inherited autosomal monosomy, or for the X chromosome in a male (XY) fetus.

[1046] *Three Chromosomes Present*

[1047] Trisomic chromosomes have three characteristic patterns. The first pattern indicates maternally-inherited meiotic trisomy, a meiotic error where the fetus inherited two homologous, non-identical chromosomes from the mother (FIG. 30E); this pattern includes two central green bands with two each of the peripheral red and blue bands. The second pattern indicates paternally-inherited meiotic trisomy, where the fetus inherited two homologous, non-identical chromosomes from the father (FIG. 30F); this pattern includes four central green bands and three each of the peripheral red and blue bands. The third pattern indicates either maternally- (FIG. 30G) or paternally-inherited (FIG. 30H) mitotic trisomy, a mitotic error where the fetus inherited two identical chromosomes from either the mother or the father; this pattern includes four central green bands with two each of the peripheral red and blue bands. Maternally- and paternally-inherited mitotic trisomies can be distinguished by the placement of the flanking red and blue bands, such that the red and blue internal peripheral bands (those not associated with the limits of the plots) are closer to the center in paternally-inherited mitotic trisomy. This is due to the paternal contribution of identical chromosomes. Note that our previous results indicate that at the blastomere stage, 66.7% of maternally-inherited trisomies are meiotic, and that only 10.2% of trisomies are paternally-inherited.

[1048] For the Y chromosome, the PS method considers a different set of hypotheses: zero, one, or two chromosomes present. As there is no maternal contribution to the sequence reads at each locus and because heterozygous loci are not possible (cases of two Y chromosomes necessarily involve two identical chromosomes), the bands remain tightly associated with the top (A alleles) or the bottom (B alleles) of the plot (data not shown), and analysis is greatly simplified, relying on quantitative allele count data. Note that since the method interrogates SNPs, it uses homologous non-recombinant SNPs from the Y chromosome, thus obtaining data on both X and Y for one probe pair.

[1049] *Identifying Aneuploidies*

[1050] Identification of autosomal aneuploidies using this plot-based visualization method is straightforward given a sufficient fetal fraction, and requires only identifying plots for which there are an abnormal number of chromosomes present, as described above. Combining the knowledge of copy number of the X and Y chromosomes identifies whether sex chromosome aneuploidies are present. Specifically, plots representing a fetus with a 47,XXX genotype will have a typical “three-chromosome” pattern, and plots representing a fetus with a 47,XXY genotype will have the typical “two-chromosome” pattern for the X chromosome, but will also have allele reads indicating the presence of one Y chromosome. The method is similarly able to call 47,XYY, where a “one chromosome” pattern indicates the presence of a single X chromosome, and allele reads indicate the presence of two Y chromosomes. A fetus with a 45,X genotype will have the typical “one-chromosome” pattern for the X chromosome, and data indicating zero Y chromosomes.

[1051] *Effects of Fetal Fraction*

[1052] As discussed above, the number of sequence reads from the fetus contributes to the precise location of each spot along the y-axis in the plots. As fetal fraction will affect the proportion of reads that originate from the fetus and the mother, it will also affect the positioning of each spot. At a high fraction of fetal cfDNA (generally above ~20%), as in FIGs. 30C-30E and FIGs. 30G and 30H, it is readily apparent that although the spots cluster based mainly on maternal genotype, the presence of fetal DNA from alleles whose genotype is distinct from the maternal genotype shift the clusters into multiple, distinct bands. However, as the fetal fraction decreases (as in FIGs. 30B and 30F), the spots regress towards the poles and center of the plot,

resulting in tighter clusters. Specifically, the set of peripheral red bands, where the maternal genotype is AA, regress towards the top of the plot; the set of peripheral blue bands, where the maternal genotype is BB, regress towards the bottom; the set of central green bands, where the mother is heterozygous, condense into a single cluster at the center of the plot (compare FIGs. 30B and 30C). Although aneuploidy is not readily apparent by eye using this visualization technique for low fetal fraction cases, the algorithm is able to identify ploidy states with a very low fetal fraction, such as 3% fetal fraction. It is able to do this because the statistical technique compares the observed data to very precise data models that predict the allele distributions for a given sample parameter set (including copy number, parental genotypes, and fetal fraction, for example). Data model precision is critical in low fetal fraction cases, as the differences between the allele distributions for different ploidy states are proportional to the fetal fraction. In addition, the algorithm is able to determine when a data set does not contain enough data to make a confident fetal ploidy determination.

[1053] *Results*

[1054] Sequencing reads that mapped to targeted SNPs were deemed to be informative and were used by the algorithm. More than 95% of targeted loci were observed in the sequencing results. The plots for visualizing key ploidy calls are depicted in FIG. 31A-31G. FIG. 31A indicates a euploid sample. Here, chromosomes 13, 18, and 21 have the typical “two chromosome” pattern (as described herein). This includes a trio of center green bands, and two red and two blue peripheral bands. This, together with the two center green bands for the X chromosome and the presence of Y chromosome bands along the plots’ peripheries, indicate a euploid XY genotype.

[1055] The most prevalent autosomal trisomies, T13, T18, and T21, are indicated by the plots in FIGs. 31B, 31C, and 31D, respectively. Specifically, FIG. 31B depicts a T13 sample. Here, chromosomes 18 and 21 display the typical “two chromosome” pattern, chromosome X displays the typical “one chromosome” pattern, and there are reads from the Y chromosome. Together, this indicates disomy at chromosomes 18 and 21, and identifies a fetal XY genotype. However, chromosome 13 depicts a typical “three chromosome” pattern – specifically. Similarly, FIG. 31C depicts a T18 sample, and FIG. 31D depicts a T21 sample.

[1056] The method is also able to detect sex chromosome aneuploidies, including 45,X (FIG. 31E), 47,XXY (FIG. 31F), and 47,XYY (FIG. 31G). Note that the method is calling copy number at chromosomes 13, 18, 21, X, and Y; the overall chromosome number is reported

assuming disomy at the remaining chromosomes. The X chromosome regions of the plot depicting a 45,X sample reveals the presence of a single chromosome. However, the lack of reads from the Y chromosome, coupled with the “two chromosome” pattern for chromosomes 13, 18, and 21, indicate a 45,X genotype. Conversely, the 47,XXY samples generate a plot revealing the presence of two X chromosomes. The data also revealed reads for alleles from the Y chromosome. Together with the presence of two copies of chromosomes 13, 18, and 21, this indicates a 47,XXY genotype. A 47,XYY genotype is indicated by the presence of a “one chromosome” pattern for the X chromosome, and reads indicating the presence of two Y chromosomes.

[1057] *Discussion*

[1058] This method detected T13, T18, T21, 45,X, 47,XXY, and 47,XYY non-invasively from maternal blood. This method interrogates cfDNA from maternal plasma by targeted multiplex PCR amplification and high-throughput sequencing of 19,488 SNPs. This, coupled with the method’s sophisticated informatics analyses that take into account parental genotypic information and numerous sample parameters, including fetal fraction and DNA quality, more robustly detects the fetal signal and makes highly accurate ploidy calls at all of the five chromosomes implicated in the seven most common types of at-birth aneuploidy (T13, T18, T21, 45,X, 47,XXX, 47,XXY, and 47,XYY). This method offers a number of clinical advantages over previous methods, including and most significantly greater clinical coverage and sample-specific calculated accuracies (analogous to a personalized risk score).

[1059] *Increased Clinical Coverage*

[1060] This method offers approximately a two-fold increase in aneuploidy coverage compared to clinically available NIPT methodologies, given its ability to accurately detect autosomal trisomies and sex chromosome aneuploidies. The method presented here is the only noninvasive test that calls ploidy at the sex chromosomes with high accuracy. Prior DNA mixing experiments and separate plasma samples analyzed in our experimental assays suggest that this method will detect a larger cohort of sex chromosome anomalies, including 47,XXX. The method presented here also detects aneuploidies at chromosomes 13, 18, and 21 with high sensitivities and

specificities, and with appropriate primer design is expected to be able to detect copy number at the remaining chromosomes as well.

[1061] *Sample-Specific Calculated Accuracies*

[1062] Significantly, this method calculates a sample-specific accuracy for ploidy calls on each chromosome in each sample. Accuracies calculated by this method are expected to significantly lower the rate of incorrect calls by identifying and flagging individual samples that have poor quality DNA or low fetal fractions that are likely to result in a poor accuracy test result. By contrast, massively parallel shotgun sequencing (MPSS)-based methods produce a positive or negative call using a single-hypothesis rejection test, and their accuracy estimate is based on a published study cohort rather than on the characteristics of the individual sample, which are assumed to have the same accuracy as the cohort. However, individual accuracies for samples with parameters in the tail of the cohort distribution may differ significantly. This is exacerbated at low fetal fractions, as in early gestational age, or for samples with low DNA quality. These samples are generally not identified and flagged for follow-up, which can result in missed calls. The present method, however, takes into account many parameters, including fetal fraction and a number of DNA quality metrics, to make each chromosome copy number call, calculating a sample-specific accuracy for that call. This allows the method to identify individual samples with low accuracy and flag them for follow-up. This is expected to nearly eliminate missed calls, especially at the early stages of pregnancy when fetal fractions are typically low. The presumption is that a no call is much preferred to a missed call, since a no call simply requires a redraw and reanalysis.

[1063] *Converting Calculated Accuracies to Traditional Risk Scores*

[1064] This method can offer an adjusted risk of aneuploidy for high-risk pregnant women, where the adjusted risk takes into account an *a priori* risk (Benn P, Cuckle H, Pergament E. Non-invasive prenatal diagnosis for Down syndrome: the paradigm will shift, but slowly. Ultrasound Obstet Gynecol 2012;39:127-130, which is hereby incorporated by reference in its entirety). Although the present method offers each patient a customized calculated accuracy, for clinical use these accuracies can be converted to traditional risk scores, which also denote the risk of an aneuploid pregnancy but are expressed as fractions. Traditional risk scores take into

account various parameters, including maternal age-related risk and serum levels of biochemical markers, to offer a risk score above which a mother is considered high-risk and for whom follow-up invasive diagnostic procedures are recommended. This method significantly refines this risk score, thus reducing both the false positive and false negative rates, and offering a more accurate assessment of individual maternal risk. A calculated accuracy as used here is the likelihood that the ploidy call is correct, and is expressed as a percentage, but the calculated accuracies used in Experiment 19 do not include an age-related risk. Because calculation of a risk score typically includes an age-related risk, the calculated accuracies and traditional risk scores are not interchangeable; they must be combined to convert into a traditional risk score. The formula to combine the age-related risk with the calculated accuracy is:

$$[1065] \frac{R_1 R_2}{R_1 R_2 + [1 - R_1][1 - R_2]}$$

[1066] where R_1 is the risk score as calculated by the present method and R_2 is the risk score as calculated by first trimester screening.

[1067] *SNP-Based Methods Negate Issues with Amplification Variation*

[1068] An inherent drawback to the counting methods used by some other methods is that they determine fetal ploidy state by measuring the ratio of the number of reads mapping to the chromosome of interest (*e.g.*, chromosome 21) to those mapping to a reference chromosome. Chromosomes with high or low GC content, including chromosomes 13, X, and Y, amplify with high variability. This can result in signal variation that is comparable in magnitude to the fetal cfDNA signal, which can confound copy number calls by altering the ratio of allele reads from the chromosome-of-interest to those from the reference chromosome. This can result in low accuracy for chromosomes 13, X, and Y. Significantly, this problem is exacerbated at low fetal cfDNA fractions, as tends to be the case at early gestational ages.

[1069] In contrast, SNP-based methods do not rely on consistent amplification levels between chromosomes, and are thus expected to provide results that are equally accurate across all chromosomes. Because the present method looks, in part, at relative counts of different alleles at polymorphic loci, which by definition differ only by a single nucleotide, it does not require the use of reference chromosomes, and this obviates the problems with chromosome-to-chromosome amplification variation that are inherent to methods that rely on quantitating read counts. Unlike quantitative methods that require reference chromosomes that are euploid, the

present method is expected to be able to detect triploidy as well as copy-number neutral anomalies like uniparental disomy.

[1070] *The Importance of Early Detection*

[1071] Significantly, the combined at-birth prevalence of sex chromosome aneuploidies is higher than that of the most common autosomal aneuploidies (FIG. 32). However, there are currently no routine non-invasive screening methods that reliably detect sex chromosome abnormalities. Thus, sex chromosome anomalies are generally detected prenatally as a side-effect of routine testing for Down syndrome or other autosomal aneuploidies; a large proportion of cases are missed entirely. Early and accurate detection is crucial for many of these disorders where early therapeutic intervention improves clinical outcomes. For example, Turner syndrome is often not diagnosed until adolescence, although its overall at-birth prevalence is 1 in 2,500 females. Growth hormone therapy is known to prevent short stature that results from the disorder, but treatments are significantly more effective when initiated prior to the age of 4. Additionally, estrogen replacement therapy can stimulate secondary sexual characteristics in patients with Turner syndrome, but again therapy must be initiated in the pre-teen years, before the syndrome is usually detected. Together, this underscores the importance of early, routine, and safe detection of sex chromosome aneuploidies. This method offers the first approach with the potential to serve as a routine screen for sex chromosome anomalies.

[1072] *Example 20*

[1073] The following experiment illustrates an exemplary method for designing and selecting primers that can be used in any of the multiplexed PCR methods of the invention. In some embodiments, primers from an initial library of candidate primers are selected so that they can be used to simultaneously amplify a large number of target loci (or a subset of target loci) in a single reaction. In some embodiments, primers from an initial library of candidate primers are selected to form multiple primer pools such that each pool can be used to simultaneously amplify a subset of target loci in a single reaction. Preferably, primers are designed and selected for a large portion or all of the most desirable target loci. Preferably, the minimum number of pools needed to amplify the target loci are created.

[1074] *Step 1*

[1075] Calculate a first score for each primer pair design using one or more of the following parameters: number of SNPs within the primers, location of SNPs within the primers, distance from an end of the amplicon to the target bases within the amplicon, number of target loci in an amplicon, heterozygosity rate of the target locus, disease prevalence associated with a sequence (*e.g.*, a polymorphism) at the target locus, disease penetrance associated with a sequence (*e.g.*, a polymorphism) at the target locus, specificity of the candidate primer for the target locus, size of the candidate primer, melting temperature of the candidate primer, melting temperature of the target amplicon, GC content of the target amplicon, GC content of the 3' end of the candidate primer, homopolymer length in the candidate primer, amplification efficiency of the target amplicon, and size of the target amplicon.

[1076] *Step 2*

[1077] Compare each primer pair to every other primer pair, and calculate a second score for the pair using one or more of the following parameters: likelihood of dimer formation, amplicon overlap, number of primer designs for a particular target locus, and distance between amplicons. In some embodiments, the score is infinite if amplicons overlap so that two different primer pairs that generate overlapping amplicons are not included in the same primer pool.

[1078] *Step 3*

[1079] Aggregate the first score and the second score together (such as by using a weighted average of the scores).

[1080] *Step 4*

[1081] If desired, order all target loci into one contiguous list based upon their genomic location in ascending order.

[1082] *Step 5*

[1083] Build a minimum priority queue data structure that prioritizes the pairs of designs (in which each design is one primer pair so that a pair of designs includes two primer pairs with a total of 4 primers) based on their score (such as the aggregate score from step 3). In some embodiments, the score for a pair of designs is the worse score (such as the worse aggregate score from step 3) out of the scores for all 4 primers in the pair of designs. The pair of designs with the best (most desirable) score is first in the queue, and the pair of designs with the worst (least desirable) score is last in the queue. If desired, pairs of designs with a score above a threshold (least desirable) are removed from the library of candidate primers such that they are not included in the final pool(s) (for example, these primers may be omitted from the queue). In some embodiments, pairs of design with an interaction score above (worse than) 20, 18, 16, 14, 12, 10, 9, 8, 7, 6, 5, 4, 3, 2, or 1 kcal/mol are removed from the library of candidate primers. In some embodiments, pairs of design with a ΔG value below (worse than) -20, -18, -16, -14, -12, -10, -9, -8, -7, -6, -5, -4, -3, -2, or -1 kcal/mol are removed from the library of candidate primers.

[1084] Each design pair can be stored as a node of a doubly linked list with initial “next” and “previous” pointers set to NULL.

[1085] *Step 6*

[1086] Begin forming all pools simultaneously by doing the following steps. Take the design pair with the best (most desirable) score from the priority queue and add it to “the potential pools.” Begin storing designs in N number of doubly linked list data structures with the design pairs. N represents the current number of different primer pools. Initially, N = 1, since there is only one primer pool. In some embodiments, a second pool is only created if necessary to include the desired target loci or the desired level of coverage of target loci. Check to see if the design pair removed from the queue is “connected” to any other existing design pair. By “connected” for purposes of this step is meant that a single design in one pair is the same as a single design in another pair. If two pairs are connected, then assign the appropriate next and previous pointers to one another. If two pairs are not connected, then add them to the “potential pools” In some embodiments, a design pair is only placed in a particular pool if it would be connected to at most two other design pairs in that pool (otherwise it can be assigned to a different pool).

[1087] Check to see if (i) any linked list spans from the first target to the last target (such that all the desired target loci are included) or (ii) if a pool meets the cutoff for the desired minimum pool level. If it does, that list now forms a pool and can be added to the “final pools” list.

[1088] *Step 7*

[1089] If desired, check to see if the desired level of coverage (such as all the bases in the target loci being included in amplicons from 4 different primer pairs) that is desired for each location. Repeat step 6 until achieving the desired level of coverage.

[1090] The resulting primer pool(s) can be used in any of the methods of the invention.

[1091] *Example 21*

[1092] The following Example illustrates an exemplary method for designing and selecting primers that can be used in any of the multiplexed PCR methods of the invention. In some embodiments, the primers are divided into different pools (e.g., 2, 3, 4, 5, 6, or more different pools) such that each pool is used to amplify target loci in a different reaction volume. Each pool is used to simultaneously amplify a large number of target loci (or a subset of target loci) in a single reaction volume. Preferably, primers are designed and selected for a large portion of the most desirable target loci or for all of the target loci. A set of candidate target loci can be selected as described in Examples 16 or 20 based on the particular polymorphisms or mutations of interest. In some embodiments, one or more of the following type of target loci are included: SNPs, short indels, long indels, exons, and combinations thereof. In some embodiments for target loci that are short indels, the PCR primer or primer pair targets a sequence of adjacent base pairs; and the indel is completely covered by one sequencing read. In some embodiments for target loci that are large indels, two primer pairs are used to target a pair of breakpoints at the boundaries of the indel. In this case, the two primer pairs are designed such that when the deletion is present there is a PCR product and the two primer pairs are selected together for inclusion in the same pool (the four primers are treated by the algorithm as a single assay rather than two assays). In some embodiments for target loci that are exons, a set of primers pairs are designed to tile the full exon.

[1093] For each candidate locus, one or more PCR primer pairs are designed using the Primer3 program (available at the worldwide web at primer3.sourceforge.net; libprimer3 release 2.2.3, which is hereby incorporated by reference in its entirety). If there are no feasible designs for PCR primers for a particular target locus, then that target locus is eliminated from further consideration. In some embodiments, each target base is covered by at least two independent PCR assays (such as two independent primer pairs that will amplify the target base) and preferably by four assays, although not all of the available assays for a target must be used. In some embodiments, no targets are omitted. Desirably, the algorithm produces as few pools as possible but may produce more than one pool. In some embodiments, two different primer pairs that are in close proximity in the genome (such as within 2 kbases or 1 kbase) and whose forward primers are on the same strand are not be assigned to the same pool. This constraint avoids primer interference in the extension-and-ligation amplification method. In some embodiments in which the PCR will be performed using a polymerase with low 5'→3' exonuclease and/or low strand displacement activity, different primer pairs that are in close proximity in the genome and whose forward primers are on the same strand can be assigned to the same pool since the with low 5'→3' exonuclease and/or low strand displacement activity of the polymerase will reduce or prevent primer interference and allow nearby or adjacent amplicons to be produced.

[1094] *Step 1*

[1095] Build an interaction graph. Each node represents one assay (such as one primer pair). Each edge represents a conflict between two assays. There are three types. Interaction edges represent a potential primer dimer and have a score indicating the interaction strength. Proximity edges represent physical proximity of the primer binding sites which may result in interference. Target edges represent redundant designs associated with the same target (a special case of a proximity edge).

[1096] *Step 2*

[1097] Select an initial value for the maximum interaction score (*e.g.*, 95% of the maximum score).

[1098] *Step 3*

[1099] Compute a score such as a utility score for each assay as follows using steps 3A and 3B.

[1100] *Step 3A*

[1101] Calculate a score for each assay based on one or more of its intrinsic characteristics. For example, favor assays with amplicons close to the optimal length (such as 300 bp); favor assays with a shorter distance from the beginning of the amplicon to the target; and/or penalize assays with primers overlapping known SNPs. Any other parameter, such as the parameters disclosed herein can also be included.

[1102] *Step 3B*

[1103] Multiply the score for each assay by a factor that varies from 0 to 1 according to the current coverage of the assay's target bases. This factor gives lower weight to targets that are already covered by assays. At the beginning of the algorithm this factor is 1 for all assays because none have been covered. Calculate the factor as follows. For each base in the target, compute a coverage score as $1 / (2^c)$ where c is the number of previously-selected assays (in other pools) that cover that base. For instance, if three assays cover the base then the coverage score is $1/(2^3) = 0.125$. The factor for the target is the maximum value of the coverage score for all bases in the target. For instance, if the target contains 10 bases, 3 bases are covered by 1 target, and 7 bases are covered by 3 targets, then the factor is $\text{MAX}(1/(2^1), 1/(2^3)) = 0.5$. The score in step 3A is then multiplied by this factor.

[1104] *Step 4*

[1105] Use a single iteration of the algorithm in Example 16 to design a pool given the current maximum interaction score: Construct a new graph with the assays that have not been assigned to a pool yet and with the edges that have weights exceeding the maximum interaction score. Remove nodes (assays) according to the algorithm in Example 16 until there are no edges left. The assay utility scores come from step 3 in this Example rather than the calculation used for Example 16.

[1106] *Step 5*

[1107] Save the assays selected in step 4 as a new pool and remove them from consideration. Then repeat steps 3 and 4 with the remaining assays, and iterate until all targets have sufficient coverage.

[1108] *Step 6*

[1109] If desired, evaluate the result. If the total number of pools meets the design goal then reduce the maximum interaction score; otherwise increase the maximum interaction score. Then go back to step 3. Iterate, using a binary search strategy to find the lowest maximum interaction score that produces the desired number of pools.

[1110] *Step 7*

[1111] Output the pools from the final iteration. After the selection process, the primers remaining in the pools may be used in any of the methods of the invention.

[1112] *Example 22*

[1113] The following Example illustrates an exemplary method for designing and selecting primers that can be used in any of the multiplexed PCR methods of the invention. In some embodiments, the primers are divided into different pools (e.g., 2, 3, 4, 5, 6, or more different pools) such that each pool is used to amplify target loci in a different reaction volume. Any of the embodiments listed in Example 21 can be used for this Example as well.

[1114] This method uses a graph coloring algorithm.

[1115] *Step 1*

[1116] Select 2, 3 or 4 of the best assays (such as primer pairs) for each target locus from all of the available assays.

[1117] *Step 2*

[1118] Select an initial maximum interaction score.

[1119] *Step 3*

[1120] Build an interaction graph containing only edges that exceed the maximum interaction score.

[1121] *Step 4*

[1122] Color the graph such that no adjacent nodes have the same color (this is a standard problem with many heuristic solutions). Each color represents a different pool.

[1123] *Step 5*

[1124] Go back to step 3 and iterate, refining the maximum interaction score until the desired number of pools is achieved. In some embodiments, after the primers are selected in step 1, the algorithm assumes all assays must be included in a pool.

[1125] After the primers are divided into different pools, the pools may be used in any of the methods of the invention.

[1126] *Example 23*

[1127] This example illustrates three exemplary methods for calculating the limit of detection for any of the methods of the invention. These methods were used to calculate the limit of detection for single nucleotide variants (SNVs) in a tumor biopsy (FIG. 38) and a plasma sample (FIG. 39).

[1128] The first method (denoted “LOD-mr5” in FIGs. 38 and 39) calculates the limit of detection based on a minimum of 5 reads being chosen as the minimum number of times a SNV is observed in the sequencing data to have sufficient confidence the SNV is actually present. The limit of detection is based on whether the observed the depth of read (DOR) is above this minimum of 5. The gray lines in FIGs. 38 and 39 indicate SNVs for which the limit of detection is limited by the DOR. In these cases, not enough reads were measured to reach the error limit of the assay. If desired, the limit of detection can be improved (resulting in a lower numerical value) for these SNVs by increasing the DOR.

[1129] The second method (denoted “LOD-zs5.0” in FIGs. 38 and 39) calculates the limit of detection based on the z-score. The Z-score is the number of standard deviations an observed

error percentage is away from the background mean error. If desired, outliers can be removed and the z-score can be recalculated and this process can be repeated. The final weighted mean and the standard deviation of the error rate are used to calculate the z-score. The mean is weighted by the DOR since the accuracy is higher when the DOR is higher.

[1130] For the exemplary z-score calculation used for this example, the background mean error and standard deviation were calculated from all the other samples of the same sequencing run weighted by their depth of read, for each genomic loci and substitution type. Samples were not considered in the background distribution if they were 5 standard deviations away from the background mean. The orange lines in FIGs. 38 and 39 indicate SNVs for which the limit of detection is limited by the error rate. For these SNV's enough reads were taken to reach the 5 read minimum, and the limit of detection was limited by the error rate. If desired, the limit of detection can be improved by optimizing the assay to reduce the error rate.

[1131] The third method (denoted "LOD-zs5.0-mr5" in FIGs. 38 and 39) calculates the limit of detection based on the maximum value of the above two metrics.

[1132] For the analysis of a tumor sample shown in FIG. 38, the mean limit of detection was 0.36%, and the median limit of detection was 0.28%. The number of DOR limited (gray lines) SNVs was 934. The number of error rate limited (orange lines) SNVs was 738.

[1133] For the analysis of cDNA in a plasma sample shown in FIG. 39, the mean limit of detection was 0.24%, and the median limit of detection was 0.09%. The number of DOR limited (gray lines) SNVs was 732. The number of error rate limited (orange lines) SNVs was 921.

[1134] *Example 24*

[1135] This example illustrates the detection of CNVs and SNVs from the same single cell. The following primer libraries were used a library of ~28,000 primers for detecting CNVs, a library of ~3,000 primers for detecting CNVs, and library of primers for detecting SNVs. For analysis of a single cell, cells were serial diluted until there were 3 or 4 cells per droplet. An individual cell was pipetted and placed into a PCR tube. The cell was lysed using Protease K, salt, and DTT using the following thermocycling conditions: 56°C for 20 minutes, 95°C for 10 minutes, and then a 4°C hold. For analysis of genomic DNA, DNA from the same cell line as the analyzed single cell was either purchased or obtained by growing the cells and extracting the DNA.

[1136] For amplification with the library of ~28,000 primers, the following PCR conditions were used a 40uL reaction volume, 7.5nM of each primer, and 2x master mix (MM). In some

embodiments QIAGEN Multiplex PCR Kit is used for the master mix (QIAGEN catalog No. 206143; *see, e.g.,* information available at the world wide web at qiagen.com/products/catalog/assay-technologies/end-point-pcr-and-rt-pcr-reagents/qiagen-multiplex-pcr-kit, is which is hereby incorporated by reference in its entirety). The kit includes 2x QIAGEN Multiplex PCR Master Mix (providing a final concentration of 3 mM MgCl₂, 3 x 0.85 ml), 5x Q-Solution (1 x 2.0 ml), and RNase-Free Water (2 x 1.7 ml). The QIAGEN Multiplex PCR Master Mix (MM) contains a combination of KCl and (NH₄)₂SO₄ as well as the PCR additive, Factor MP, which increases the local concentration of primers at the template. Factor MP stabilizes specifically bound primers, allowing efficient primer extension by, *e.g.,* HotStarTaq DNA Polymerase. HotStarTaq DNA Polymerase is a modified form of *Taq* DNA polymerase and has no polymerase activity at ambient temperatures. The following thermocycling conditions were used for the first round of PCR: 95°C for 10 minutes; 25 cycles of 96°C for 30 seconds, 65°C for 29 minutes, and 72°C for 30 seconds; and then 72°C for 2 minutes, and a 4°C hold. For the second round of PCR a 10ul reaction volume, 1xMM, and 5nM of each primer was used. The following thermocycling conditions were used: 95°C for 15 minutes; 25 cycles of 94°C for 30 seconds, 65°C for 1 minute, 60°C for 5 minutes, 65°C for 5 minutes, and 72°C for 30 seconds; and then 72°C for 2 minutes, and a 4°C hold.

[1137] For the library of ~3,000 primers, exemplary reaction conditions include a 10ul reaction volume, 2x MM, 70mM TMAC, and 2nM primer of each primer. For the library of primers for detecting SNVs, exemplary reaction conditions include a 10ul reaction volume, 2x MM, 4mM EDTA, and 7.5nM primer of each primer. Exemplary thermocycling conditions include 95°C for 15 minutes, 20 cycles of 94°C for 30 seconds, 65°C for 15 minutes, and 72°C for 30 seconds; and then 72°C for 2 minutes, and a 4°C hold.

[1138] The amplified products were barcoded. One run of sequencing was performed with an approximately equal number of reads per sample.

[1139] FIGs. 40A and 40B show results from analysis of genomic DNA (FIG. 40A) or DNA from a single cell (FIG. 40B) using a library of approximately 28,000 primers designed to detect CNVs. Approximately 4 million reads were measured per sample. The presence of two central bands instead of one central band indicates the presence of a CNV. For three samples of DNA from a single cell, the percent of mapped reads was 89.9%, 94.0%, and 93.4%, respectively. For two samples of genomic DNA the percent of mapped reads was 99.1% for each sample.

[1140] FIGs. 41A and 41B show results from analysis of genomic DNA (FIG. 41A) or DNA from a single cell (FIG. 41B) using a library of approximately 3,000 primers designed to detect CNVs. Approximately 1.2 million reads were measured per sample. The presence of two central bands instead of one central band indicates the presence of a CNV. For three samples of DNA from a single cell, the percent of mapped reads was 98.2%, 98.2%, and 97.9%, respectively. For two samples of genomic DNA the percent of mapped reads was 98.8% for each sample. FIG. 42 illustrates the uniformity in DOR for these ~3,000 loci.

[1141] For calling SNVs, the call percent for true positive mutations was similar for DNA from a single cell and genomic DNA. A graph of call percent for true positive mutations for single cells on the y-axis versus that for genomic DNA on the x-axis yielded a curve fit of $y = 1.0076x - 0.3088$ with $R^2 = 0.9834$. FIG. 43 shows similar error call metrics for genomic DNA and DNA from a single cell. FIG. 44 shows that the error rate for detecting transition mutations was greater than for detecting transversion mutations, indicating it may be desirable to select transversion mutations for detection rather than transition mutations when possible. In some embodiments, at least 50, 60, 70, 80, 90, 95, 96, 98, 99, or 100% of the SNVs tested for are transversion mutations rather than transition mutations.

[1142] *Example 25*

[1143] The following is an example of multiplex PCR conditions in which the annealing temperature is significantly higher than the average or maximum melting (T_m) of the primers in the library. A 3,168-plex reaction was performed with 3,168 primer pairs to 3,168 different target loci. For the PCR amplification a 20 ul total volume was used with 2 nM of each primer (3,168 pairs of forward and reverse primers), 70 mM TMAC (tetra-methyl ammonium chloride), and 7 ul library DNA or genomic DNA. The following thermocycling conditions were used: 95°C for 10 minutes and then 25 cycles of 96°C for 30 seconds, 65°C for 20 minutes (this annealing temperature is higher than the T_m of the primers, listed above), and 72°C for 30 seconds. Then, 72°C for 2 minutes and a 4°C hold were used.

[1144] The minimum T_m (the lowest numerical value for the T_m for any of the primers) for this primer library is 54.0 °C. The maximum T_m (the highest numerical value for the T_m for any of the primers) for this primer library is 60.36 °C. The average T_m (average value of the T_m values of the primers) for this primer library is 55.25 °C. These T_m values were calculated using the following exemplary method for calculating T_m values. This method is used by the Primer3

program (the worldwide web at primer3.sourceforge.net, which is hereby incorporated by reference in its entirety) to calculate T_m values. In some embodiments, one or more of the following conditions are assumed for this calculation: temperature: of 60.0 °C, primer concentration of 100 nM, and/or salt concentration of 100 mM. In some embodiments, other conditions are assumed for this calculation, such as the conditions that will be used for multiplex PCR with the library.

[1145] $T_m = \Delta H / (\Delta S + R \ln(C/4))$

[1146] Below is documentation from the Primer3 program for its T_m calculations; PRIMER_TM_FORMULA (int; default 0) specifies details of melting temperature calculation. This is new in version 1.1.0, and added by Maido Remm and Triinu Koressaar (the world wide web at primer3.ut.ee/primer3web_help.htm#PRIMER_TM_FORMULA, which is hereby incorporated by reference in its entirety). A value of 0 directs primer3 to a backward compatible calculation (in other words, the only calculation available in previous version of primer3). This backward compatible calculation uses the table of thermodynamic parameters in the paper (Breslauer KJ *et al.* (1986) "Predicting DNA duplex stability from the base sequence" *Proc Natl Acad Sci* 83:4746-50, [dx.doi.org/10.1073/pnas.83.11.3746](https://doi.org/10.1073/pnas.83.11.3746), which is hereby incorporated by reference in its entirety), and the method in the paper (Rychlik W, Spencer WJ and Rhoads RE (1990) "Optimization of the annealing temperature for DNA amplification in vitro", *Nucleic Acids Res* 18:6409-12, [dx.doi.org/10.1093/nar/18.21.6409](https://doi.org/10.1093/nar/18.21.6409), which is hereby incorporated by reference in its entirety).

[1147] A value of 1 (which is recommended) directs primer3 to use the table of thermodynamic values and the method for melting temperature calculation suggested in the following paper (SantaLucia JR (1998) "A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics", *Proc Natl Acad Sci* 95:1460-65, [dx.doi.org/10.1073/pnas.95.4.1460](https://doi.org/10.1073/pnas.95.4.1460), which is hereby incorporated by reference in its entirety). The tag PRIMER_SALT_CORRECTIONS can be used to specify the salt correction method for melting temperature calculation.

[1148] The following is an example of calculating the melting temperature of an oligo with PRIMER_TM_FORMULA=1 and PRIMER_SALT_CORRECTIONS=1 recommended values for primer=CGTGACGTGACGGACT.

[1149] Using default salt and DNA concentrations gives

$$[1150] \quad T_m = \Delta H / (\Delta S + R \cdot \ln(C/4))$$

[1151] where R is the gas constant (1.987 cal/K mol) and C is the DNA concentration.

$$[1152] \quad \Delta H(\text{predicted}) = \Delta H(\text{CG}) + \Delta H(\text{GT}) + \Delta H(\text{TG}) + \dots + \Delta H(\text{CT}) + \Delta H(\text{init.w.term.GC}) + \Delta H(\text{init.w.term.AT})$$

$$[1153] \quad = -10.6 + (-8.4) + (-8.5) + \dots + (-7.8) + 0.1 + 2.3$$

$$[1154] \quad = -128.8 \text{ kcal/mol}$$

[1155] where 'init.w.term GC' and 'init.w.term AT' are two initiation parameters for duplex formation: 'initiation with terminal GC' and 'initiation with terminal AT.'

$$[1156] \quad \Delta S(\text{predicted}) = \Delta S(\text{CG}) + \Delta S(\text{GT}) + \Delta S(\text{TG}) + \dots + \Delta S(\text{CT}) + \Delta S(\text{init.w.term.GC}) + \Delta S(\text{init.w.term.AT})$$

$$[1157] \quad = -27.2 + (-22.4) + (-22.7) + \dots + (-21.0) + (-2.8) + 4.1$$

$$[1158] \quad = -345.2 \text{ cal/k}^*\text{mol}$$

$$[1159] \quad \Delta S(\text{salt corrected}) = \Delta S(\text{predicted}) + 0.368 \cdot 15(\text{NN pairs}) \cdot \ln(0.05\text{M monovalent cations})$$

$$[1160] \quad = -361.736$$

$$[1161] \quad T_m = -128.800 / (-361.736 + 1.987 \cdot \ln((5 \cdot 10^{-(8)})/4)) = 323.704 \text{ K}$$

$$[1162] \quad T_m(C) = 323.704 - 273.15 = 50.554 \text{ C}$$

[1163] *Additional Applications*

[1164] Because this method utilizes targeted amplification, it is uniquely poised to detect sub-microscopic anomalies, such as microdeletions and microduplications. Although non-targeted methods like MPSS have been shown to detect the DiGeorge microdeletion syndrome, this required a sufficiently high level of genomic coverage so as to make the approach unfeasible. This is because non-targeted amplification will be several orders of magnitude less efficient on

submicroscopic regions, as very small fraction of the sequencing reads will be informative. Additionally, the fact that the currently available methods have trouble accurately identifying ploidy state for the sex chromosomes suggests that they will also encounter variable amplification problems on smaller chromosomal segments.

[1165] Similarly, SNP based methods can detect UPD disorders, which are copy number-neutral anomalies that will not be detected by either the current noninvasive methods that rely on counting or the traditional invasive methods like amniocentesis and CVS that rely on cytogenetic karyotyping and/or fluorescence *in situ* hybridization. This is because SNP-based methods are uniquely able to distinguish individual haplotypes, whereas the clinically available MPSS-based and targeted methods amplify non-polymorphic loci and are thus unable to determine, for example, whether the chromosomes-of-interest originate from the same parent. This means that these microdeletion/microduplication and UPD syndromes, including Prader-Willi, Angelman, and Beckwith-Wiedemann syndromes, are generally not diagnosed prenatally, and are often initially misdiagnosed postnatally. This significantly delays therapeutic intervention. Additionally, because this method targets SNPs, this method will also facilitate parental haplotype reconstruction, allowing for detection of fetal inheritance of individual disease-linked loci (Kitzman JO, Snyder MW, Ventura M, *et al.* Noninvasive whole-genome sequencing of a human fetus. *Sci Transl Med* 2012;4:137ra76, which is hereby incorporated by reference in its entirety).

[1166] The results presented here confirm the expanded scope of this method for identifying prenatal aneuploidy. Specifically, by amplifying and sequencing 19,488 SNPs, this method is able to determine copy number at chromosomes 13, 18, 21, X, and Y, and is uniquely expected to detect other chromosomal abnormalities, such as triploidy and UPD, that are not detected by any other clinically available non-invasive method. The increased clinical coverage and powerful sample-specific calculated accuracies suggest that this method may offer a viable adjunct to invasive testing for detecting fetal chromosomal aneuploidies.

[1167] *Example 26*

[1168] This example describes an exemplary method for detection of copy number variations in breast cancer samples using SNP-targeted massively multiplexed PCR. Evaluation of CNV in tumor tissues typically involves SNP microarray or aCGH. These methods have high whole-

genome resolution, but require large amounts of input material, have high fixed costs, and do not work well on formaldehyde fixed-paraffin embedded (FFPE) samples. For this example, 28,000-plex SNP-targeted PCR with next generation sequencing (NGS) was used to target 1p, 1q, 2p, 2q, 4p16, 5p15, 7q11, 15q, 17p, 22q11, 22q13 and chromosomes 13, 18, 21 and X for detection of CNVs in breast cancer samples. Accuracy was validated on 96 samples with aneuploidies or microdeletions. Single-molecule sensitivity was established by analyzing single cells. Of 17 breast cancer samples (15 fresh frozen and 2 FFPE tumor tissues, 5 pairs of matched tumor and normal cell lines) analyzed, 16 (including both FFPEs) were observed with full or partial CNVs in one to 15 targets (average: 7.8); evidence of tumor heterogeneity was observed. The three tissues with one CNV all had a 1q duplication, the most frequent cytogenetic abnormality in breast carcinoma. The most frequent regions with CNVs were 1q, 7p, and 22q1. Only one tumor tissue (with 9 CNVs) had a region with LOH; this LOH was also detected in adjacent putatively normal tissue that lacked the other 8 CNVs. By contrast, 5 or more regions with LOH and a high total CNV incidence (average: 12.8) was detected in cell lines. Thus, massively multiplexed PCR offers an economical high-throughput approach to investigate CNVs in a targeted manner, and is applicable to difficult-to-analyze samples, such as FFPE tissues.

[1169] *Example 27*

[1170] This example further validates a massively multiplexed PCR methodology for chromosomal aneuploidy and CNV determination disclosed herein, sometimes referred to as CoNVERGe (Copy Number Variant Events Revealed Genotypically), in cancer diagnostics, and further illustrates the development and use of “PlasmArt” standards for PCR of ctDNA samples. PlasmArt standards include polynucleotides having sequence identity to regions of the genome known to exhibit CNV and a size distribution that reflects that of cfDNA fragments naturally found in plasma.

[1171] *Sample Collection*

[1172] Human breast cancer cell lines (HCC38, HCC1143, HCC1395, HCC1937, HCC1954, and HCC2218) and matched normal cell lines (HCC38BL, HCC1143BL, HCC1395BL, HCC1937BL, HCC1954BL, and HCC2218BL) were obtained from the American Type Culture Collection (ATCC). Trisomy 21 B-lymphocyte (AG16777) and paired father / child DiGeorge Syndrome (DGS) cell lines (GM10383 and GM10382, respectively) were from the Coriell Cell Repository (Camden, NJ). GM10382 cells only have the paternal 22q11.2 region.

[1173] We procured tumour tissues from 16 breast cancer patients, including 11 fresh frozen (FF) samples from Geneticist (Glendale, CA) and five formalin-fixed paraffin-embedded (FFPE) samples from North Shore-LIJ (Manhasset, NY). We acquired matched buffy coat samples for eight patients and matched plasma samples for nine patients. FF tumour tissues and matched buffy coat and plasma samples from five ovarian cancer patients were from North Shore-LIJ. For eight breast tumour FF samples, tissue subsections were resected for analysis. Institutional review board approvals from Northshore/LIJ IRB and Kharkiv National Medical University Ethics Committee were obtained for sample collection and informed consent was obtained from all subjects.

[1174] Blood samples were collected into EDTA tubes. Circulating cell free DNA (containing ctDNA) was isolated from 1 mL plasma using the QIAamp Circulating Nucleic Acid Kit (Qiagen, Valencia, CA). Genomic DNA (gDNA) from FF tumor tissues, blood, and buccal samples was extracted using the DNeasy Blood and Tissue Kit (Qiagen).

[1175] To make the PlasmArt standards according to one exemplary method, first, 9×10^6 cells were lysed with hypotonic lysis buffer (20 mM Tris-Cl (pH 7.5), 10 mM NaCl, and 3 mM $MgCl_2$) for 15 min on ice. Then, 10% IGEPAL CA-630 (Sigma, St. Louis, MO) was added to a final concentration of 0.5%. After centrifugation at 3,000 g for 10 min at 4°C, pelleted nuclei were resuspended in 1X micrococcal nuclease (MNase) Buffer (New England BioLabs, Ipswich, MA) before adding 1000 U of MNase (New England BioLabs), and then incubated for 5 min at 37°C. Reactions were stopped by adding EDTA to a final concentration of 15 mM. Undigested chromatin was removed by centrifugation at 2,000 g for 1 min. Fragmented DNA was purified with the DNA Clean & Concentrator™-500 kit (Zymo Research, Irvine, CA). Mononucleosomal DNA produced by MNase digestion was also purified and size-selected using AMPure XP magnetic beads (Beckman Coulter, Brea, CA). DNA fragments were sized and quantified with a Bioanalyzer DNA 1000 chip (Agilent, Santa Clara, CA).

[1176] To model ctDNA at different concentrations, different fractions of PlasmArts from HCC1954 and HCC2218 cancer cells were mixed with those from the corresponding matched normal cell line (HCC1954BL and HCC2218BL, respectively). Three samples at each concentration were analyzed. Similarly, to model allelic imbalances in plasma DNA in a focal 3.5 Mb region, we generated PlasmArts from DNA mixtures containing different ratios of DNA from a child with a maternal 22q11.2 deletion and DNA from the father. Samples containing only the father's DNA were used as negative controls. Eight samples at each concentration were analyzed.

[1177] *Massively multiplexed PCR and DNA sequencing*

[1178] Massively multiplex PCR and DNA sequencing methods below were used to determine allele counts at a plurality of polymorphic loci with 3–6 million (M) reads/sample for cell lines, 1.5–7 M reads/sample for tumour tissues, 18 M reads/sample for FFPE-LCM samples, 6–7 M reads/sample for germline controls, and 18–25 M reads/sample for plasma. For two representative exemplary runs using the 3,168 SNP primer pair pool, an average of 20 million reads were used to obtain allele counts for plasma DNA libraries and 6 million reads were used to obtain allele counts for genomic DNA libraries from fresh-frozen human tumors. The percent of mapped reads (i.e. mapped to the human genome) on these two exemplary runs were 98% and 95%, respectively. The fraction of sequencing reads at a given locus with a particular allele (allele fraction) was the fractional abundance of the allele in a sample. These counts provided observed allele frequencies that were used by the data analysis methods provided immediately below in this Example to determine the ploidy state of a chromosome or chromosome segment of interest and/or to determine the average allelic imbalance of the sample.

[1179] Libraries were generated from the samples above. Adapters were ligated to DNA fragments and the fragments were amplified using the following protocol: 95°C, 2 min; 15 x [95°C, 20 sec, 55°C, 20 sec, 68°C, 20 sec], 68°C 2 min, 4°C hold.

[1180] Multiplexed PCR allows simultaneous amplification of many targets in a single reaction. In this study, we targeted 3,168 SNPs, which were distributed across five chromosome arms as follows: 646 on 1p, 602 on 1q, 541 on 2p, 707 on 2q, and 672 on the 22q11.2 focal region. These genomic regions were selected for convenience from SNP panels available in our laboratory. Target SNPs had at least 10% population minor allele frequency (1000 Genomes Project data; April 30, 2012 release) to ensure that a sufficient fraction would be heterozygous in any given patient. For each SNP, multiple primers were designed to have a maximum amplicon length of 75 bp and a melting temperature between 54.0–60.5°C. To minimize the likelihood of primer dimer product formation, primer interaction scores for all possible combinations of primers were calculated, and primers with high scores were eliminated. The 3,186 SNP primer pair pool all had ΔG values greater than -4Kcal/mol. Candidate PCR assays were ranked and 3,168 assays were selected on the basis of target SNP minor-allele frequency, observed heterozygosity rate (from dbSNP), presence in HapMap, and amplicon length.

[1181] For PCR amplifications, 3,168 SNPs were amplified in a multiplex PCR reaction using one primer pair for each SNP, during 25 cycles, and sequencing barcodes were added in 12 additional cycles. Prior to sequencing, the barcoded products were pooled, purified with the QIAquick PCR Purification Kit (Qiagen), and quantified using the Qubit™ dsDNA BR Assay Kit (Life Technologies). Amplicons were sequenced using an Illumina HiSeq 2500 sequencer with 1.5–7 M reads/sample for tumor tissue DNA and 18–25 M reads/sample for plasma cfDNA.

[1182] For the 3,168 SNP multiplex PCR reaction, approximately 7 ul (approx. 1200 ng) of library DNA, such as DNA from a DNA library generated from plasma of a target individual, was used. The master mix included the following: 2x (twice manufacturer's recommended concentration) Qiagen master mix, 70 mM TMAC (tetramethylammonium chloride, Sigma), 2 nM each primer, and 7 ul nucleic acid library (~1200 ng total library input) (20 ul total volume). The cycling conditions for the 3,168 SNP multiplex PCR reaction were as follows: 95°C, 15 min; 25 x [96°C, 30 sec; 65°C, 20 min; 72°C, 30 sec]; 72°C, 2 min; 4°C hold.

[1183] For the barcoding reaction, a 1X master mix was prepared that included the following: 1 uM forward primer (containing Illumina sequencing tag), 1 uM reverse primer (containing Illumina sequencing tag as well as internally-designed sequencing barcode), 1 ul of mmPCR product, diluted 1:2,000, and 1x Qiagen master mix. Barcoding cycling conditions were as follows: 95°C, 10 min; 12 x [95°C, 30 sec; 70°C, 10 sec, 60°C, 30 sec; 65°C, 15 sec, 72°C, 15 sec]; 72°C, 2 min; 4°C hold.

[1184] *Data Analysis of Tumor Tissue Genomic DNA*

[1185] For tumor tissue samples, CNVs were delineated by transitions between allele frequency distributions. Regions with at least 100 SNPs that had an allele ratio statistically different from 0.50 were considered to be of interest. More specifically, the analysis focused on regions with average allele ratios of ≤ 0.45 or ≥ 0.55 for loci that are heterozygous in the germline. A segmentation algorithm was used to exhaustively search DNA sequences in five chromosome arms as follows: 646 on 1p, 602 on 1q, 541 on 2p, 707 on 2q, and 672 on the 22q11.2 for such regions, and iteratively selected them starting from the longest one until a region of 100 SNPs was reached. Once a ≥ 100 SNP region was determined to contain a CNV, it was further segmented by average allelic ratios with a minimum segment size of 50 SNPs if needed.

[1186] Fresh frozen tissue samples from three patients with breast cancer were also analyzed using Illumina CytoSNP-12 microarrays as previously described (Levy, B. et al. Genomic imbalance in products of conception: single-nucleotide polymorphism chromosomal microarray analysis. *Obstetrics and gynecology* 124, 202-209 (2014)).

[1187] Data Analysis of Circulating Tumor DNA

[1188] CNVs were identified by a maximum likelihood algorithm that searched for plasma CNVs in regions where the tumor sample from the same individual also had CNVs, using haplotype information deduced from the tumor sample. This algorithm modeled expected allelic frequencies across a set of average allelic imbalances at 0.025% intervals for three sets of hypotheses: (1) all cells are normal (no allelic imbalance), (2) some/all cells have a homolog 1 deletion or homolog 2 amplification, or (3) some/all cells have a homolog 2 deletion or homolog 1 amplification. For at least some of the analysis, modeling was performed up to 15% average allelic imbalance, although for the vast majority of samples AAI was less than or equal to 5%. The likelihood of each hypothesis was determined at each SNP using a Bayesian classifier based on a beta binomial model of expected and observed allele frequencies at all heterozygous SNPs, and then the joint likelihood across multiple SNPs was calculated taking linkage of the SNP loci into consideration. The maximum likelihood hypothesis from the comparison of expected to observed allele frequencies was then selected.

[1189] This algorithm also calculates the confidence of each CNV call by comparing the likelihoods of different hypotheses. A confidence threshold of 99.9% was used in plasma samples to minimize false positive results.

[1190] For dimorphic SNPs that have alleles arbitrarily designated 'A' and 'B', the allele ratio of the A allele is $n_A/(n_A + n_B)$, where n_A and n_B are the number of sequencing reads for alleles A and B, respectively. Allelic imbalance is the difference between the allele ratios of A and B for loci that are heterozygous in the germline. This definition is analogous to that for SNVs, where the proportion of abnormal DNA is typically measured using mutant allele frequency, or $n_m/(n_m + n_r)$, where n_m and n_r are the number of sequencing reads for the mutant allele and the reference allele, respectively.

[1191] Allele frequency data was corrected for errors before it was used to generate individual probabilities. Errors that were corrected included allele amplification bias, ambient contamination, genotype contamination, and sequencing error. Ambient contamination refers to the contamination error across all SNPs in addition to sequencing errors, and genotype contamination refers to the additional contamination at some SNPs due to contamination from another sample. Ambient contamination and genotype contamination were determined on the same run as the on-test sample

analysis by analyzing homozygous alleles in the sample. The ploidy status of a chromosomal segment was estimated using heterozygous loci for a test individual.

[1192] Best hypothesis was defined to be the one with the highest likelihood across all polymorphic loci. Likelihood at each locus was calculated using a beta binomial model of observed allele frequencies at each of the polymorphic loci, and the likelihood across a set of polymorphic loci was computed using the phase information deduced from the corresponding tumor sample.

[1193] A linear regression model was used to compare either expected AAI or tumor input DNA percentage and observed AAI determined by the CNV detection algorithm. $P < 0.05$ was considered statistically significant. SigmaPlot 12.5 (Systat Software, San Jose, CA) and Matlab 7.12.0 R2011.a (MathWorks, Natick, MA) were used.

[1194] Accordingly, to evaluate the sensitivity and reproducibility of CoNVERGe, especially when the proportion of abnormal DNA for a CNV, or average allelic imbalance (AAI), is low, we used it to detect CNVs in DNA mixtures comprised of a previously characterized abnormal sample titrated into a matched normal sample. The mixtures consisted of artificial cfDNA, termed “PlasmArt”, with fragment size distribution approximating natural cfDNA (see above). In the first pair, a son’s tumor DNA sample having a 3 Mb Focal CNV deletion of the 22q11.2 region was titrated into a matched normal sample from the father at between 0–1.5% total cfDNA. CoNVERGe reproducibly identified CNVs corresponding to the known abnormality with estimated AAI of $>0.35\%$ in mixtures of $\geq 0.5\% \pm 0.2\%$ AAI, failed to detect the CNV in 6/8 replicates at 0.25% abnormal DNA, and reported a value of $\leq 0.05\%$ for all eight negative control samples. The AAI values estimated by CoNVERGe showed high linearity ($R^2 = 0.940$) and reproducibility (error variance = 0.087). The assay was sensitive to different levels of amplification within the same sample. Based on these data a conservative detection threshold of 0.45% AAI could be used for subsequent analyses.

[1195] Two additional PlasmArt titrations, prepared from pairs of matched tumor and normal cell line samples and having CNVs on chromosome 1 or chromosome 2, were also evaluated. Among negative controls, all values were $<0.45\%$, and high linearity ($R^2 = 0.952$ for HCC1954 1p, $R^2 = 0.993$ for HCC1954 1q, $R^2 = 0.977$ for HCC2218 2p, $R^2 = 0.967$ for HCC2218 2q) and reproducibility (error variance = 0.190 for HCC1954 1p, 0.029 for HCC1954 1q, 0.250 for HCC2218 2p, and 0.350 for HCC2218 2q) were observed between the known input DNA amount and that calculated by CoNVERGe. The difference in the slopes of the regressions for regions 1p and 1q of

one sample pair correlates with the relative difference in copy number observed in the B-allelic frequencies (BAFs) of regions 1p and 1q of the same sample, demonstrating the relative precision of the AAI estimate calculated by CoNVERGe.

[1196] CoNVERGe has application to a variety of sample sources including FFPE, Fresh Frozen, Single Cell, Germline control and cfDNA. We applied CoNVERGe to six human breast cancer cell lines and matched normal cell lines to assess whether it can detect somatic CNVs. Arm-level and focal CNVs were present in all six tumour cell lines, but were absent from their matched normal cell lines, with the exception of chromosome 2 in HCC1143 in which the normal cell line exhibits a deviation from the 1:1 homolog ratio. To validate these results on a different platform, we performed CytoSNP-12 microarray analyses, which produced consistent results for all samples. Moreover, the maximum homolog ratios for CNVs identified by CoNVERGe and CytoSNP-12 microarrays exhibited a strong linear correlation ($R^2 = 0.987$, $P < 0.001$).

[1197] We next applied CoNVERGe to fresh-frozen (FF) and formalin-fixed, paraffin-embedded (FFPE) breast tumour tissue samples. In both sample types, several arm-level and focal CNVs were present; however, no CNVs were detected in DNA from matched buffy coat samples. CoNVERGe results were highly correlated with those from microarray analyses of the same samples ($R^2 = 0.909$, $P < 0.001$ for CytoSNP-12 on FF; $R^2 = 0.992$, $P < 0.001$ for OncoScan on FFPE). CoNVERGe also produces consistent results on small quantities of DNA extracted from laser capture microdissection (LCM) samples, for which microarray methods are not suitable.

[1198] *Detection of CNVs in single cells with CoNVERGe*

[1199] To test the limits of the applicability of this mmPCR approach, we isolated single cells from the six aforementioned cancer cell lines and from a B-lymphocyte cell line that had no CNVs in the target regions. The CNV profiles from these single-cell experiments were consistent between three replicates and with those from genomic DNA (gDNA) extracted from a bulk sample of about 20,000 cells. On the basis of the number of SNPs with no sequencing reads, the average assay drop-out rate for bulk samples was 0.48% (range: 0.41–0.60%), which is attributable to either synthesis or assay design failure. For single cells, the additional average assay drop-out rate observed was 0.39% (range: 0.19–0.67%). For single cell assays that did not fail (i.e. no assay drop-out occurred), the average single ADO rate calculated using heterozygous SNPs only was 0.05% (range: 0.00–0.43%). Additionally, the percentage of SNPs with high confidence genotypes (i.e. SNP genotypes determined with at least 98% confidence) was similar for both single cell and bulk samples and the genotype in the single cell samples matched those in the bulk sample (average 99.52%, range: 92.63–

100.00%).

[1200] In single cells, allele frequencies are expected to directly reflect chromosome copy numbers, unlike in tumour samples where this may be confounded by TH and non-tumour cell contamination. BAFs of $1/n$ and $(n - 1)/n$ indicate n chromosome copies in a region. Chromosome copy numbers are indicated on the allele frequency plots for both single cells and matched gDNA samples.

[1201] *Application of CoNVERGe to plasma samples*

[1202] To investigate the ability of CoNVERGe to detect CNVs in real plasma samples, we applied our approach to cfDNA paired with a matched tumour biopsy from each of two stage II breast cancer patients and five late-stage ovarian cancer. In all seven patients, CNVs were detected in both FF tumour tissues and in the corresponding plasma samples. A total of 32 CNVs, at a level of $\geq 0.45\%$ AAI, were detected in the seven plasma samples (range: 0.48–12.99% AAI) over the five regions assayed, which represent about 20% of the genome. Note that the presence of CNVs in plasma cannot be confirmed due to the lack of alternative orthogonal methods.

[1203] Although AAI estimates may appear correlated with BAFs in tumour, direct proportionality should not necessarily be expected due to tumour heterogeneity. For example, in sample BC5, regions that have BAFs compatible with $N = 11$ were detecting; combining this with the AAI calculation from the plasma sample leads to estimates for c of 2.33% and 2.67% for the two regions. Estimating c using the other regions in the sample give values between 4.46% and 9.53%, which clearly demonstrates the presence of tumor heterogeneity.

[1204] Further CNV analyses of tumor tissue samples

[1205] We applied our mmPCR-NGS method described herein to plasma samples from four stage II breast cancer patients (BC1–BC4), and analyzed the concordance between CNVs detected in plasma and those detected in multiple tumor samples from each corresponding patient. Thus, we analyzed 4–6 tissue subsections from a tumor from each of four patients with breast cancer using mmPCR-NGS. All subsections for each patient had a CNV detected in at least one of the five targeted genomic regions (1p, 1q, 2p, 2q, and 22q11.2). A CNV was identified in at least one tumor subsection in 18/20 (90%) genomic regions. Among these 18 CNV-positive regions, 11 (61%) had a CNV detected in that particular region in all subsections.

[1206] Interestingly, different patterns of AAIs across these five chromosomal regions were observed among different tumor subsections. In patient BC1, for instance, a similar pattern of CNVs was observed for regions 2p, 2q, and 22q11.2 in all four subsections, suggesting that these CNVs

are clonal mutations. In contrast, only two of the four subsections had CNVs observed in the 1p region, and three of the four subsections had CNVs observed in the 1q region, suggesting that those CNVs were subclonal mutations. Similar patterns of possible clonal and subclonal CNVs were observed in patients BC3 and BC4, whereas BC2 appeared to be more homogenous.

[1207] In addition, even when a CNV was detected in all subsections for a particular patient, such as in the 1q region for patient BC3, the AAI often varied between subsections. Overall AAI patterns also differed between patients. Taken together, these findings suggest that mmPCR-NGS can be used to elucidate both intra- and inter-tumor clonal heterogeneity.

[1208] *Concordance of CNVs in tumors and plasma cfDNA*

[1209] To quantify the amount of overlap between CNVs detected in plasma cfDNA and those detected in tumor tissue gDNA, we used mmPCR-NGS to interrogate CNVs in tumor tissue samples and matching plasma samples from patients BC1–BC4. Seven of the 18 (39%) CNV-positive genomic regions identified in tumor subsections were also detected in the plasma (0.77%–5.80% AAI). Considering only the 11 clonal CNVs—those that were detected in all tumor subsections—a CNV was detected in four (36%) of the patient-matched plasma samples (estimated AAI: 0.77%–5.80%).

[1210] Among the seven subclonal CNVs—those that were not observed in all subsections—we detected a CNV in 3/7 (43%) of the regions (estimated AAI: 1.24%–3.36%) in the corresponding cfDNA. Of note, in these three regions (BC1, chromosome 1p; BC1, chromosome 1q; and BC4, chromosome 2p), a CNV was detected in 10/14 (71%) of the matched tumor subsections. In contrast, in the other four genomic regions that did not have a CNV detected in the corresponding plasma samples (BC3, chromosomal regions 1p, 2p, 2q, and 22q11.2), we only detected a CNV in 7/24 (29%) of the tissue subsections. These data suggest that the more prevalent a subclonal CNV is within a tumor, the more likely it is to be observed in cfDNA.

[1211] In the 150 genomic regions assayed in 30 negative controls, there were no CNVs with AAIs >0.45% and confidence >99.9%, which suggests that mmPCR-NGS has a low false-positive rate.

[1212] These data demonstrate that CNVs can be detected in plasma in a substantial fraction of samples, and suggest that the more prevalent a CNV is within a tumour, the more likely it is to be observed in cfDNA. Furthermore, CoNVERGe detected CNVs from a liquid biopsy that may have otherwise gone unobserved in a traditional tumour biopsy.

[1213] Example 28

[1214] This example provides details regarding certain exemplary sample preparation methods used for analysis of different types of samples. The sample preparation methods disclosed in this example, were used in other Examples provided herein, to generate nucleic acid templates spanning a plurality of SNP sites for next generation sequencing reactions. From these NGS reactions, allele counts were generated at a plurality of polymorphic loci. These counts were then used by the analytical methods provided herein, to determine the ploidy state of a chromosome or chromosome segment of interest and/or to determine the average allelic imbalance of a sample.

[1215] *Single Cell CNV Protocol for 28,000-plex PCR*

[1216] Multiplexed PCR allows simultaneous amplification of many targets in a single reaction. Target SNPs were identified in each genomic region with 10% minimum population minor allele frequency (1000 Genomes Project data; April 30, 2012 release). For each SNP, multiple primers, semi-nested, were designed to have an amplicon length of a maximum length of 75 bp and a melting temperature between 54–60.5°C. Primer interaction scores for all possible combinations of primers were calculated; primers with high scores were eliminated to reduce the likelihood of primer dimer product formation. Candidate PCR assays were ranked and selected on the basis of target SNP minor allele frequency, observed heterozygosity rate (from dbSNP), presence in HapMap, and amplicon length.

[1217] In certain experiments, single cell samples were prepared and amplified using a mmPCR 28,000-plex protocol. The samples were prepared in the following way: For analysis of a single cell, cells were serially diluted until there were 3 or 4 cells per droplet. An individual cell was pipetted and placed into a PCR tube. The cell was lysed using Protease K, salt, and DTT using the following conditions: 56°C for 20 minutes, 95°C for 10 minutes, and then a 4°C hold. For analysis of genomic DNA, DNA from the same cell line as the analyzed single cell was either purchased or obtained by growing the cells and extracting the DNA. The DNA was amplified in a 40µL reaction volume containing Qiagen mp-PCR master mix (2XMM final conc), 7.5nM primer conc. for 28K primer pairs having hemi-nested Rev primers under the following conditions: 95C 10min, 25x [96C 30sec, 65C 29min, 72C 30sec], 72C 2min, 4C hold. The amplification product was diluted 1:200 in water and 2 µl added to STAR 2 (10 µl reaction volume) 1XMM, 5nM primer conc. and PCR was performed using hemi-nested inner Fwd primer and tag specific Rev primer: 95C 15min, 25x [94C 30sec, 65C 1min, 60C 5min, 65C 5min, 72C 30sec], 72C 2min, 4C hold.

[1218] Full sequence tags and barcodes were attached to the amplification products and amplified for 9 cycles using adaptor specific primers. Prior to sequencing, the barcoded library product were pooled, purified with the QIAquick PCR Purification Kit (Qiagen), and quantified using the Qubit dsDNA BR Assay Kit (Life Technologies). Amplicons were sequenced using an Illumina HiSeq 2500 sequencer.

[1219] *Extraction of DNA from a Blood/Plasma Sample*

[1220] Blood samples were collected into EDTA tubes. The whole blood sample was centrifuged and separated into three layers: the upper layer, 55% of the blood sample, was plasma and contains cell-free DNA (cfDNA); the buffy coat middle layer contained leucocytes having DNA, <1% of total; and the bottom layer, 45% of the collected blood sample, contained erythrocytes, no DNA was present in this fraction as erythrocytes are enucleated. Circulating tumor DNA was isolated from at least 1 mL plasma using the QIAamp Circulating Nucleic Acid Kit, Qia-Amp (Qiagen, Valencia, CA), according to the manufacture's protocol. In certain experiments genomic DNA (gDNA) from FF tumor tissues, blood, and buccal samples was extracted using the DNeasy Blood and Tissue Kit (Qiagen).

[1221] *Plasma CNV Protocol for 3,168-plex for Chromosomes 1p, 1q, 2p, 2q, and 22q11*

[1222] Plasma DNA libraries were prepared and amplified using a mmPCR 3,168-plex protocol. The samples were prepared in the following way: Up to 20 mL of blood was centrifuged to isolate the buffy coat and the plasma. Plasma extraction of cfDNA and library preparation was performed. DNA was eluted in 50 uL TE buffer. The input for mmPCR was 6.7 uL of amplified and purified Natera plasma library at an input amount of approximately 1200 ng. The plasma DNA was amplified in a 20uL reaction volume containing Qiagen mp-PCR master mix (2XMM final conc), 2 nM tagged primer conc. (total 12.7 uM) in 3,168-plex primer pools and PCR amplified: 95C 10min, 25x [96C 30sec, 65C 20min, 72C 30sec], 72C 2min, 4C hold. The amplification product was diluted 1:2,000 in water and 1 ul added to the Barcoding-PCR in a 10 uL reaction volume. The barcodes were attached to the amplification products via PCR amplification for 12 cycles using tag specific primers. Products of multiple samples were pooled and then purified with QIAquick PCR Purification Kit (Qiagen) and eluted in 50 ul DNA suspension buffer. Samples were sequenced by NGS as described for the Single Cell CNV Protocol for 28,000-plex PCR.

[1223] *Breast cancer feasibility SNV panel from plasma*

[1224] cfDNA from breast cancer patient blood samples was prepared and amplified using 336 primer pairs that were distributed into four 84-plex pools. Natera plasma libraries were prepared as

described for Plasma CNV Protocol for 3,168-plex for Chromosomes 1p, 1q, 2p, 2q, and 22q11. DNA was eluted in 50 uL TE buffer. The input for mPCR was 2.5uL of amplified and purified Natera plasma library at an input amount of approximately 600 ng. SNPs were selected from the 1000 Genomes map for Humans, Group 19 and dbSNP to pick targets, but only SNPs from the 1000 Genomes were used to screen for minor allele frequencies. The plasma DNA was amplified in four parallel reactions of 84-plex primer pools, a 10uL reaction volume containing Qiagen mp-PCR master mix (2XMM final conc.), 4mM EDTA, 7.5 nM primer concentration (total 1.26 uM) and PCR amplified: 95C 15min, 25x [94C 30sec, 65C 15min, 72C 30sec], 72C 2min, 4C hold. The amplification product of the 4 subpools were each diluted 1:200 in water and 1 ul added to the Barcoding-PCR reaction in a 10 uL reaction volume containing Q5 HS HF master mix (1xfinal), and 1 uM each barcoding primer and each of the pools were amplified in the following reaction: 98C 1min, 25x [98C 10sec, 70C 10sec, 60C 30sec, 65C 15sec, 72C 15sec], 72C 2min, 4C hold. Libraries were purified with QIAquick PCR Purification Kit (Qiagen) and eluted in 50 ul DNA suspension buffer. Samples were sequenced by paired end sequencing.

[1225] Example 29

[1226] This example demonstrates that by using low primer concentrations such that primer amount is the limiting reactant in multiplex PCR in a workflow that is followed by next generation sequencing, uniformity of density of reads, and therefore limits of detection, across a pool of amplification reactions is improved. Some experiments were carried out for plasma CNV using the 3,168-plex panel according to Example 28 above except that the total reaction volume was 10 uL instead of 20 uL. Furthermore, PCR was carried out for 15, 20, or 25 cycles. Other experiments were carried out using the four 84-plex pools on breast cancer samples according to the protocol of Example 28 except that primer concentrations were 2 nM and PCR amplification was carried out for 15, 20, or 25 cycles.

[1227] Not to be limited by theory, it is believed that primer limited multiplex PCR provides improved depth of read uniformity for multiplex PCR before multi-read sequencing, such as high throughput or massively parallel sequencing, such as sequencing on an Illumina HiSeq or MiSeq system or an Ion Torrent PGM or Proton system, based on the following considerations: If some of the amplifications in a multiplex PCR have lower efficiencies than others, then with normal multiplex PCR we will end up with a wide range of depth of read (“DOR”) values. However, if the amount of primer is limited, and the multiplex PCR is cycled more times than what it takes to exhaust

the primers, then the more efficient amplifications will stop doubling (because they have no more primers to use) and the less efficient ones will continue to double; this will result in a more similar amount of amplification product for all of the amplification products. This will translate into a much more uniform distribution of the DOR.

[1228] The following calculations are used to determine the number of cycles that would exact a given amount of primer and starting nucleic acid template:

[1229] -assume a given starting DNA input level: 100k copies of each target (10^5 ; this is easily achieved with using amplified library)

[1230] -assume we use 2 nM of each primer as an exemplary concentration, although other concentrations such as, for example, 0.2, 0.5, 1, 1.5, 2, 2.5, 5, or 10 nM could work too.

[1231] -calculate the number of primer molecules for each primer: 2×10^{-9} (molar concentration, 2 nM) $\times 10 \times 10^{-6}$ (reaction volume, 10 ul) $\times 6 \times 10^{23}$ (number of molecules per mole, Avogadro's number) = 12×10^9

[1232] -calculate the amplification fold needed to consume all primers: 12×10^9 (number of primer molecules) / 10^5 (number of copies of each target) = 12×10^4

[1233] -calculate the number of cycles needed to achieve this amplification fold, assuming 100% efficiency at each cycle: $\log_2(12 \times 10^4) = 17$ cycles. (this is $\log_2$ because at each cycle, the number of copies doubles).

[1234] So for these conditions (100k copies input, 2 nM primers, 10 ul reaction volume, assuming 100% PCR efficiency at each cycle), the primers would be consumed after 17 PCR cycles.

[1235] However, the key assumption is that some of the products DO NOT have 100% efficiency, so without measuring their efficiencies (which is only practicable for a small number of them anyway), it would take more than 17 cycles to consume them.

[1236] For each of four 84-plex SNV PCR primer pools we observed improved DOR efficiency with increasing cycles from 15 to 20 to 25. Similar results were obtained for experiments using the 3,168-plex panel. The limit of detection decreased (i.e. SNV sensitivity increased) with increasing depth of read. Furthermore, the sensitivity was consistently better when detecting transversion

mutations than transition mutations. It is likely that additional increases in DOR efficiency can be obtained with additional cycles when using primer-limiting multiplex PCR before multi-read sequencing.

[1237] Accordingly, in one aspect provided herein is a method of amplifying a plurality of target loci in a nucleic acid sample that includes (i) contacting the nucleic acid sample with a library of primers and other primer extension reaction components to provide a reaction mixture, wherein the relative amount of each primer in the reaction mixture compared to the other primer extension reaction components creates a reaction wherein the primers are present at a limiting concentration, and wherein the primers hybridize to a plurality of different target loci; and (ii) subjecting the reaction mixture to primer extension reaction conditions for sufficient number of cycles to consume or exhaust the primers in the library of primers, to produce amplified products that include target amplicons. For example, the plurality of different target loci can include at least 2, 3, 5, 10, 25, 50, 100, 200, 250, 500, 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; or 100,000 different target loci, and at most, 50, 100, 200, 250, 500, 1,000; 2,000; 5,000; 7,500; 10,000; 20,000; 25,000; 30,000; 40,000; 50,000; 75,000; 100,000, 200,000, 250,000, 500,000, and 1,000,000 different target loci to produce a reaction mixture.

[1238] The method in illustrative embodiments, includes determining an amount of primer that will be a rate limiting amount. This calculation typically includes estimating and/or determining the number of target molecules and involves analyzing and/or determining the number of amplification cycles performed. For example, in illustrative embodiments, the concentration of each primer is less than 100, 75, 50, 25, 10, 5, 2, 1, 0.5, 0.25, 0.2 or 0.1 nM. In various embodiments, the GC content of the primers is between 30 to 80%, such as between 40 to 70% or 50 to 60%, inclusive. In some embodiments, the range of GC content (*e.g.*, the maximum GC content minus minimum GC content, such as 80% - 60% = a range of 20%) of the primers is less than 30, 20, 10, or 5%. In some embodiments, the melting temperature (T_m) of the primers is between 40 to 80 °C, such as 50 to 70 °C, 55 to 65 °C, or 57 to 60.5 °C, inclusive. In some embodiments, the range of melting temperatures of the primers is less than 20, 15, 10, 5, 3, or 1 °C. In some embodiments, the length of the primers is between 15 to 100 nucleotides, such as between 15 to 75 nucleotides, 15 to 40 nucleotides, 17 to 35 nucleotides, 18 to 30 nucleotides, 20 to 65 nucleotides, inclusive. In some embodiments, the primers include a tag that is not target specific, such as a tag that forms an internal loop structure. In some embodiments, the tag is between two DNA binding regions. In various embodiments, the

primers include a 5' region that is specific for a target locus, an internal region that is not specific for the target locus and forms a loop structure, and a 3' region that is specific for the target locus. In various embodiments, the length of the 3' region is at least 7 nucleotides. In some embodiments, the length of the 3' region is between 7 and 20 nucleotides, such as between 7 to 15 nucleotides, or 7 to 10 nucleotides, inclusive. In various embodiments, the test primers include a 5' region that is not specific for a target locus (such as a tag or a universal primer binding site) followed by a region that is specific for a target locus, an internal region that is not specific for the target locus and forms a loop structure, and a 3' region that is specific for the target locus. In some embodiments, the range of the length of the primers is less than 50, 40, 30, 20, 10, or 5 nucleotides. In some embodiments, the length of the target amplicons is between 50 and 100 nucleotides, such as between 60 and 80 nucleotides, or 60 to 75 nucleotides, inclusive. In some embodiments, the range of the length of the target amplicons is less than 100, 75, 50, 25, 15, 10, or 5 nucleotides.

[1239] In various embodiments of any of the aspects of the invention, the primer extension reaction conditions are polymerase chain reaction conditions (PCR). In various embodiments, the length of the annealing step is greater than 3, 5, 8, 10, or 15 minutes but less than 240, 120, 60, or 30 minutes. In various embodiments, the length of the extension step is greater than 3, 5, 8, 10, or 15 minutes but less than 240, 120, 60 or 30 minutes.

[1240] All patents, patent applications, and published references cited herein are hereby incorporated by reference in their entirety. While the methods of the present disclosure have been described in connection with the specific embodiments thereof, it will be understood that it is capable of further modification. Furthermore, this application is intended to cover any variations, uses, or adaptations of the methods of the present disclosure, including such departures from the present disclosure as come within known or customary practice in the art to which the methods of the present disclosure pertain, and as fall within the scope of the appended claims. For example, any of the methods disclosed herein for DNA can be readily adapted for RNA by including a reverse transcription step to convert the RNA into DNA. Examples that use polymorphic loci for illustration can be readily adapted for the amplification of nonpolymorphic loci if desired. Any of the methods disclosed herein can be used with a low level of multiplexing if desired (such as with less than 1,000, 750, 500, 250, 100, 75, 50, 25, or 10 different primers or different primer pairs to different target loci).

CLAIMS

What is claimed is:

1. A method of sequencing target loci in a nucleic acid sample, the method comprising:
 - (a) contacting in one reaction volume, a nucleic acid sample comprising DNA molecules comprising target human loci and a library of at least 1,000 non-immobilized, non-identical primers that hybridize to at least 1,000 non-identical target human loci to produce a reaction mixture in solution in which the concentration of each primer is less than 20 nM; wherein the primers do not include molecular inversion probes (MIPs); and wherein the average length of the DNA molecules is less than 200 base pairs;
 - (b) subjecting the reaction mixture to primer extension reaction conditions that include an annealing step to produce amplified products comprising target amplicons; wherein the length of the annealing step for the reaction conditions is greater than 10 minutes; wherein the length of the target amplicons is less than 100 nucleotides; wherein at least 1,000 different target human loci are amplified, and wherein at least 50% of the amplified products comprise target loci; and
 - (c) sequencing the amplified human loci, wherein the method does not comprise using a microarray, thereby sequencing target loci in the nucleic acid sample.

ABSTRACT OF THE DISCLOSURE

[1241] The invention provides methods for simultaneously amplifying multiple nucleic acid regions of interest in one reaction volume as well as methods for selecting a library of primers for use in such amplification methods. The invention also provides library of primers with desirable characteristics, such as minimal formation of amplified primer dimers or other non-target amplicons.

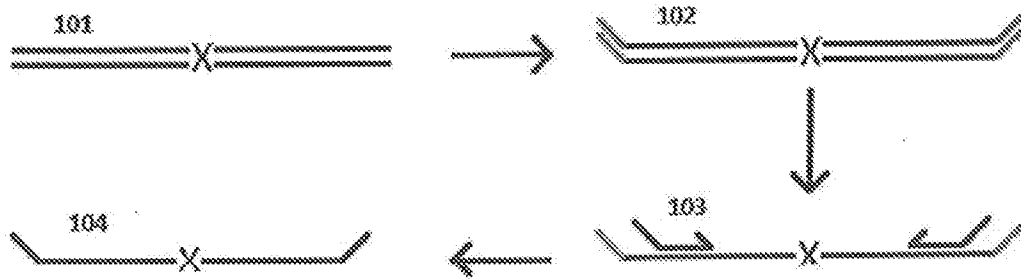


FIG. 1

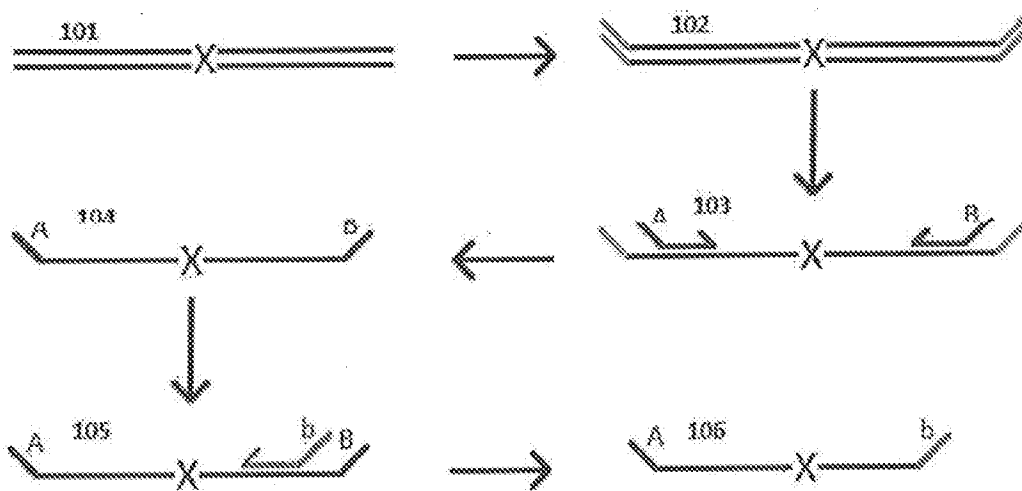


FIG. 2

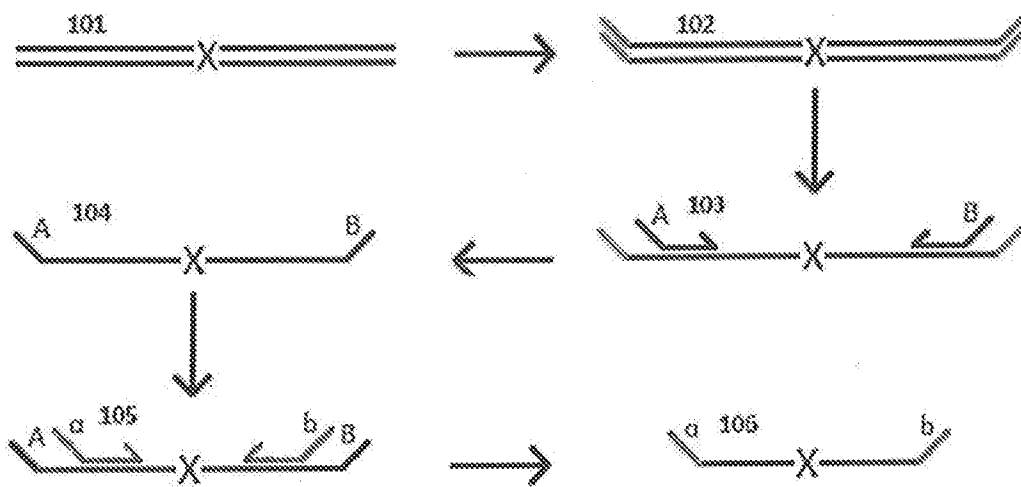


FIG. 3

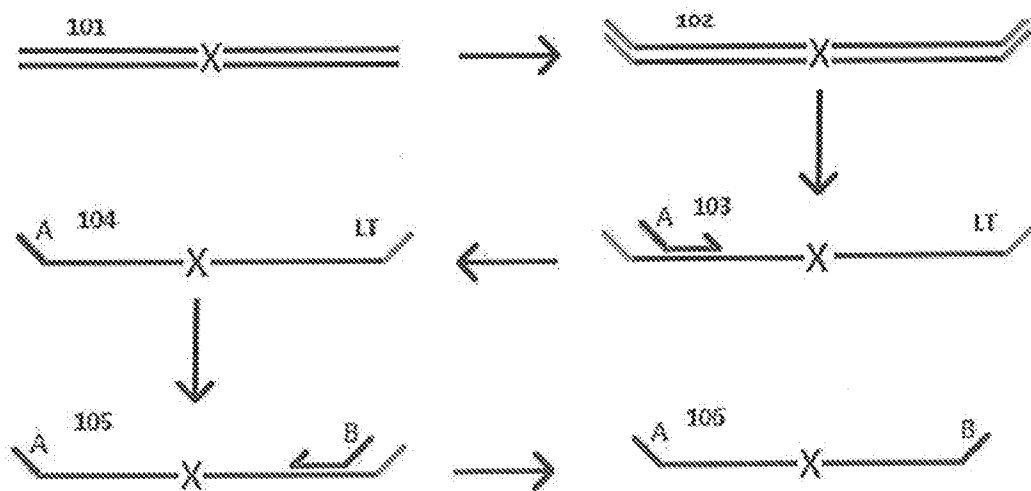


FIG. 4

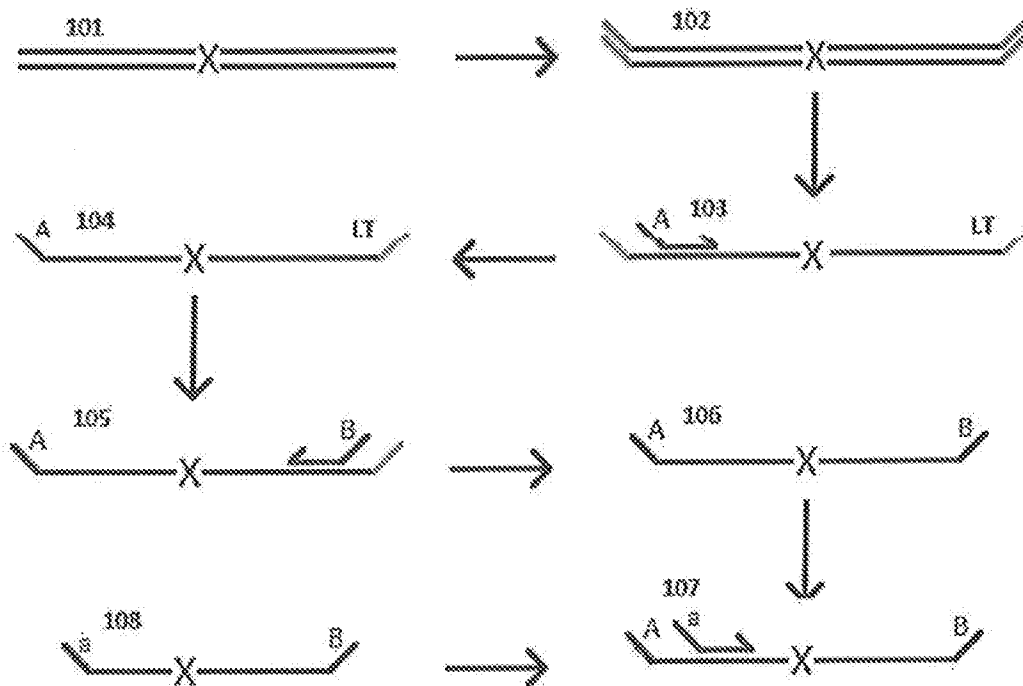


FIG. 5

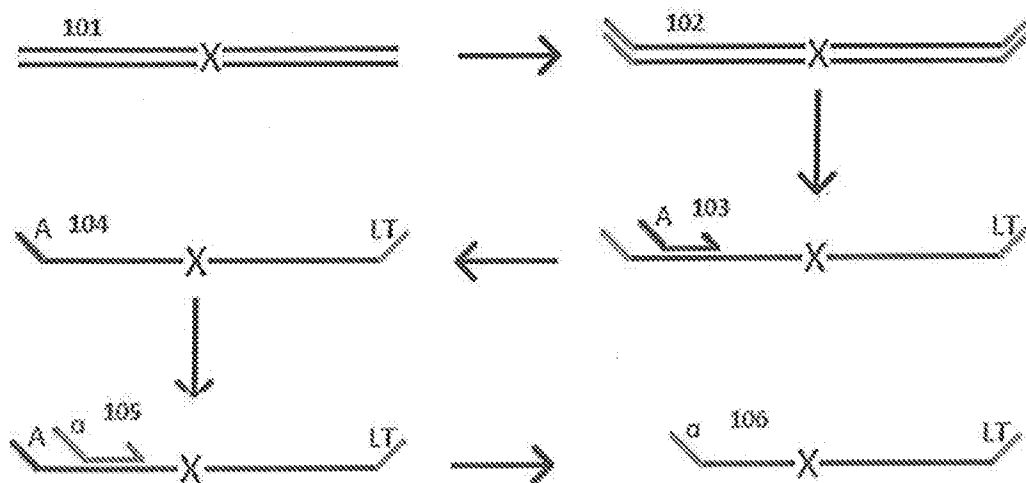


FIG. 6

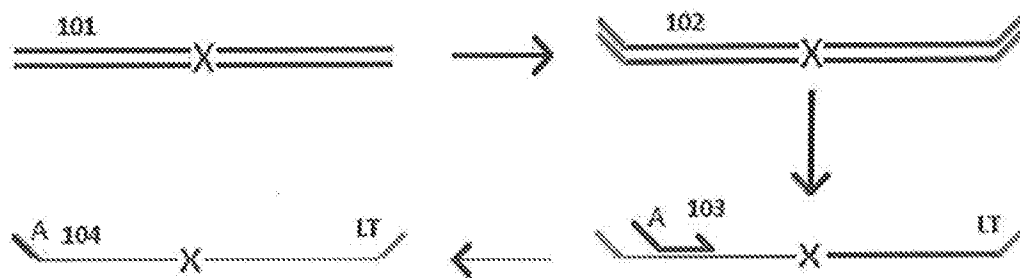


FIG. 7

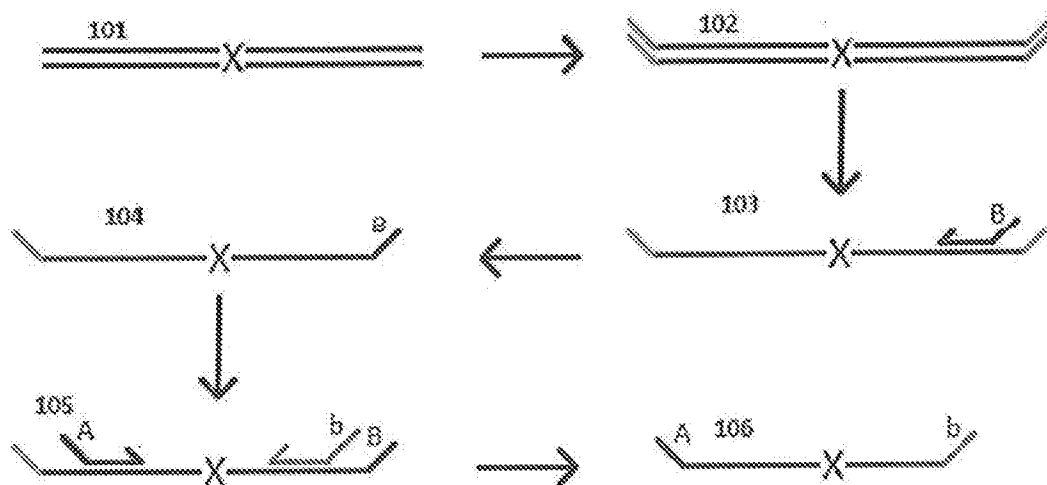


FIG. 8

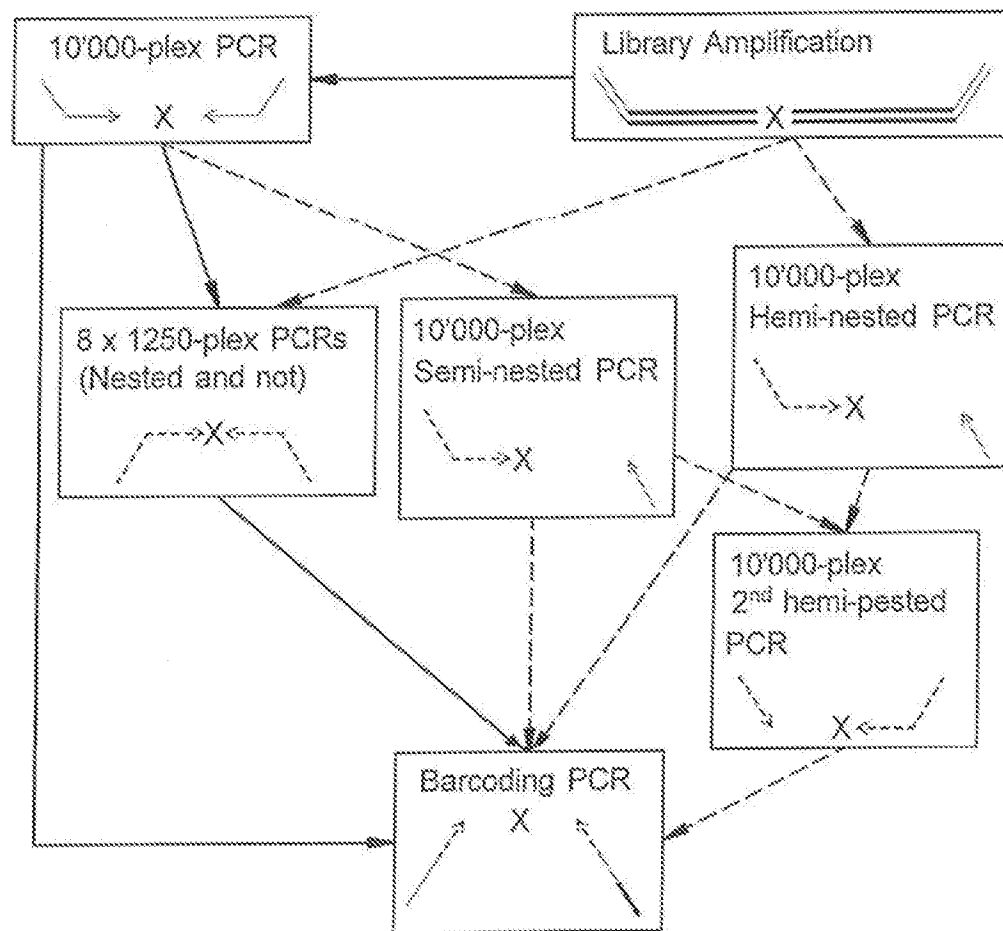


FIG. 9

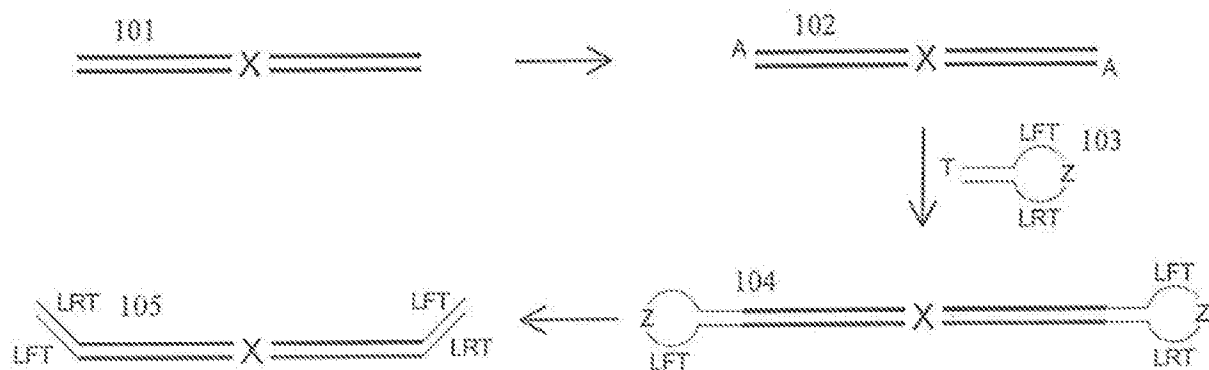


FIG. 10

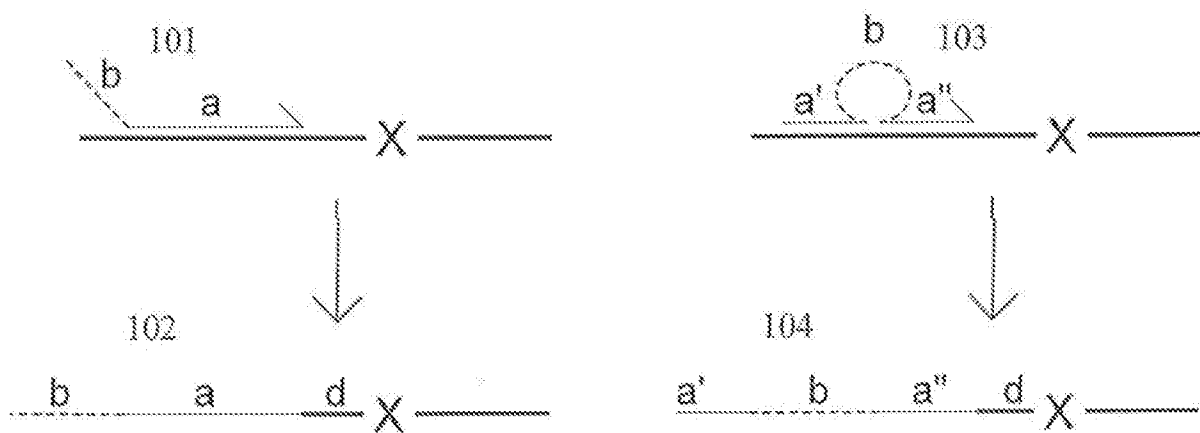


FIG. 11

The sequencing-adaptor sequence is located inside the primer sequence and flanked by target specific sequence on both sides. 10 bases are target-specific at the 3'-end of each primer. Primers were tested successfully in real-time PCR. For sequencing this reduces the number of primer bases that need to be sequenced.	
rs8130584	int-tag 1.10 AACTCACAATAGC <u>ACACGACGCTCTTCGGAATCT</u> TGCAAGCACA
rs2832093	int-tag 2.10 TCCTCTGTG <u>ACACGACGCTCTTCGGAATCT</u> CCCTGCTCTT
rs12011281	int-tag 3.10 Tcctctgtgacgg <u>ACACGACGCTCTTCGGAATCT</u> GGGCTGTCA
rs6719561	int-tag 4.10 TACATCCTTGAG <u>ACGACGCTCTTCGGAATCT</u> CGGATCTGCTGTCAGT
rs10187018	int-tag 5.10 ttgtctgagc <u>ACACGACGCTCTTCGGAATCT</u> cgggagtttc
rs10460481	int-tag 6.10 gtctatggtg <u>ACACGACGCTCTTCGGAATCT</u> caaagccagt
The sequencing-adaptor sequence is located inside the primer sequence and flanked by target specific sequence on both sides. The internal tag is formed into a hairpin structure by 10 complementary bases on either end. This brings the target-specific ends of the primer into close proximity and hinders unspecific binding to the "internal tag". 10 bases are target-specific at the 3'-end of each primer. Primers were tested successfully in real-time PCR.	
rs8130584	loop-int-tag 1.10 AACTCACAATAGC <u>ggtctggtgacacgacgctcttcggaatct</u> TGCAAGCACA
rs2832093	loop-int-tag 2.10 TCCTCTGTG <u>ggtctggtgacacgacgctcttcggaatct</u> CCCTGCTCTT
rs12011281	loop-int-tag 3.10 tctctctgtgacgg <u>gacacgacgctcttcggaatct</u> cgggctgtca
rs6719561	loop-int-tag 4.10 TACATCCTTGAG <u>gtgacgggtacacgacgctcttcggaatct</u> CGGATCTGCTGTCAGT
rs10187018	loop-int-tag 5.10 ttgtctgagc <u>gtgacgggtacacgacgctcttcggaatct</u> cgggagtttc
rs10460481	loop-int-tag 6.10 gtctatggtg <u>gtgacgggtacacgacgctcttcggaatct</u> caaagccagt

Fig. 12

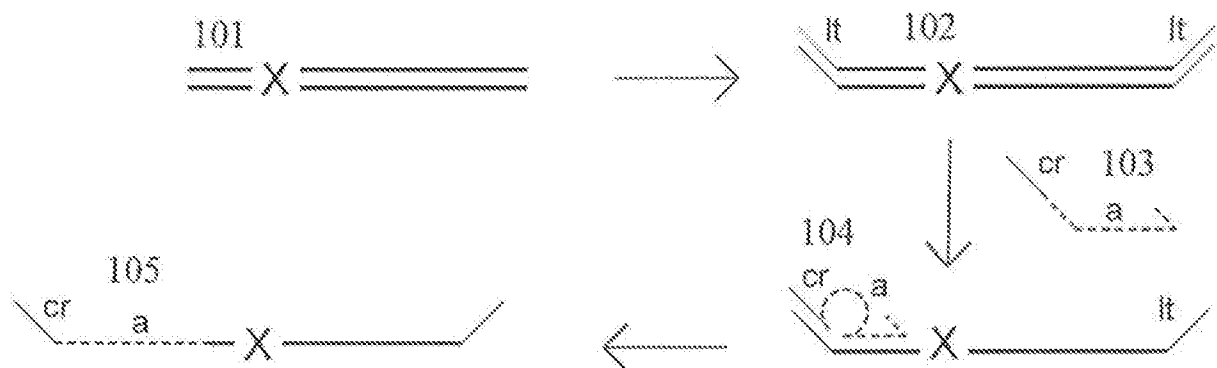


FIG. 13

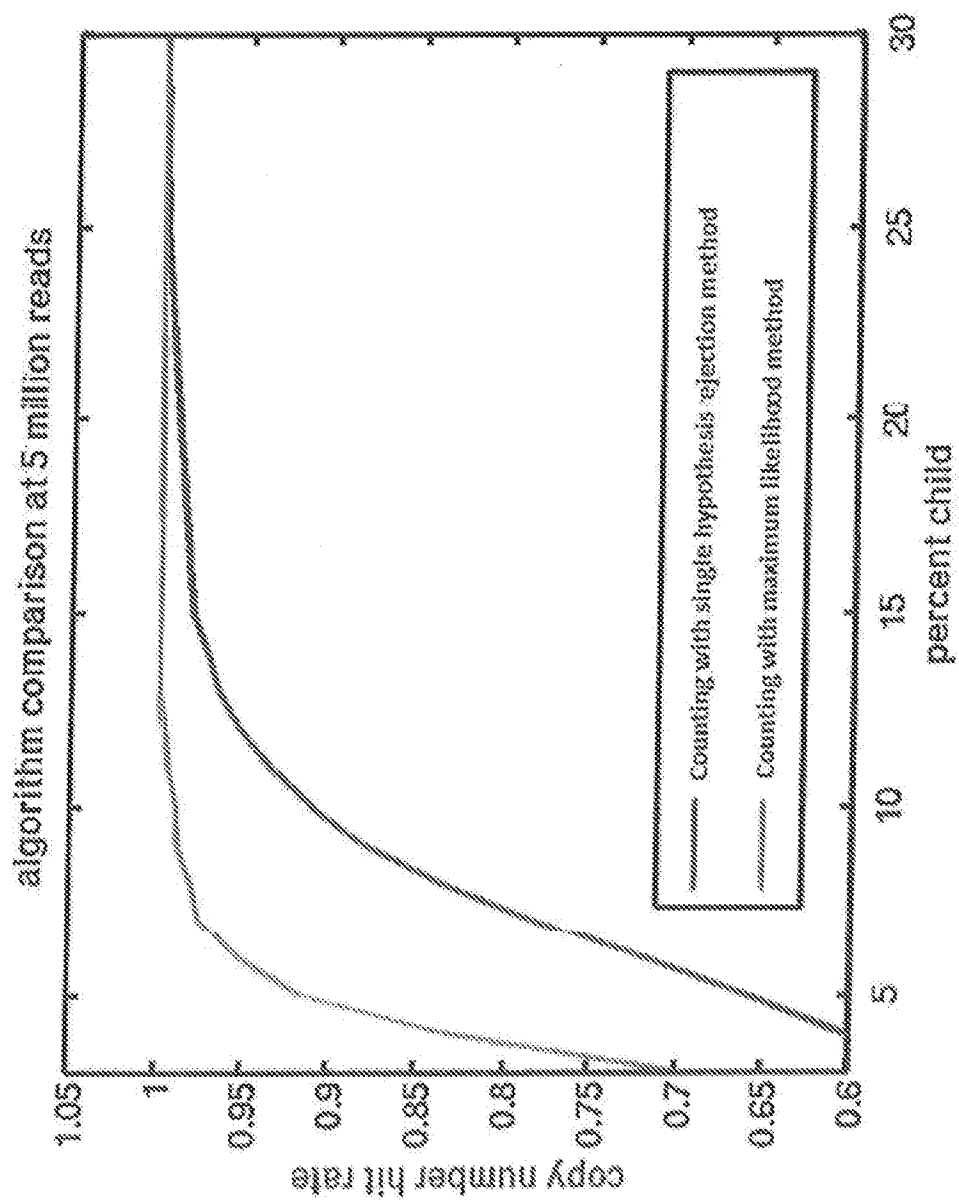


FIG. 14

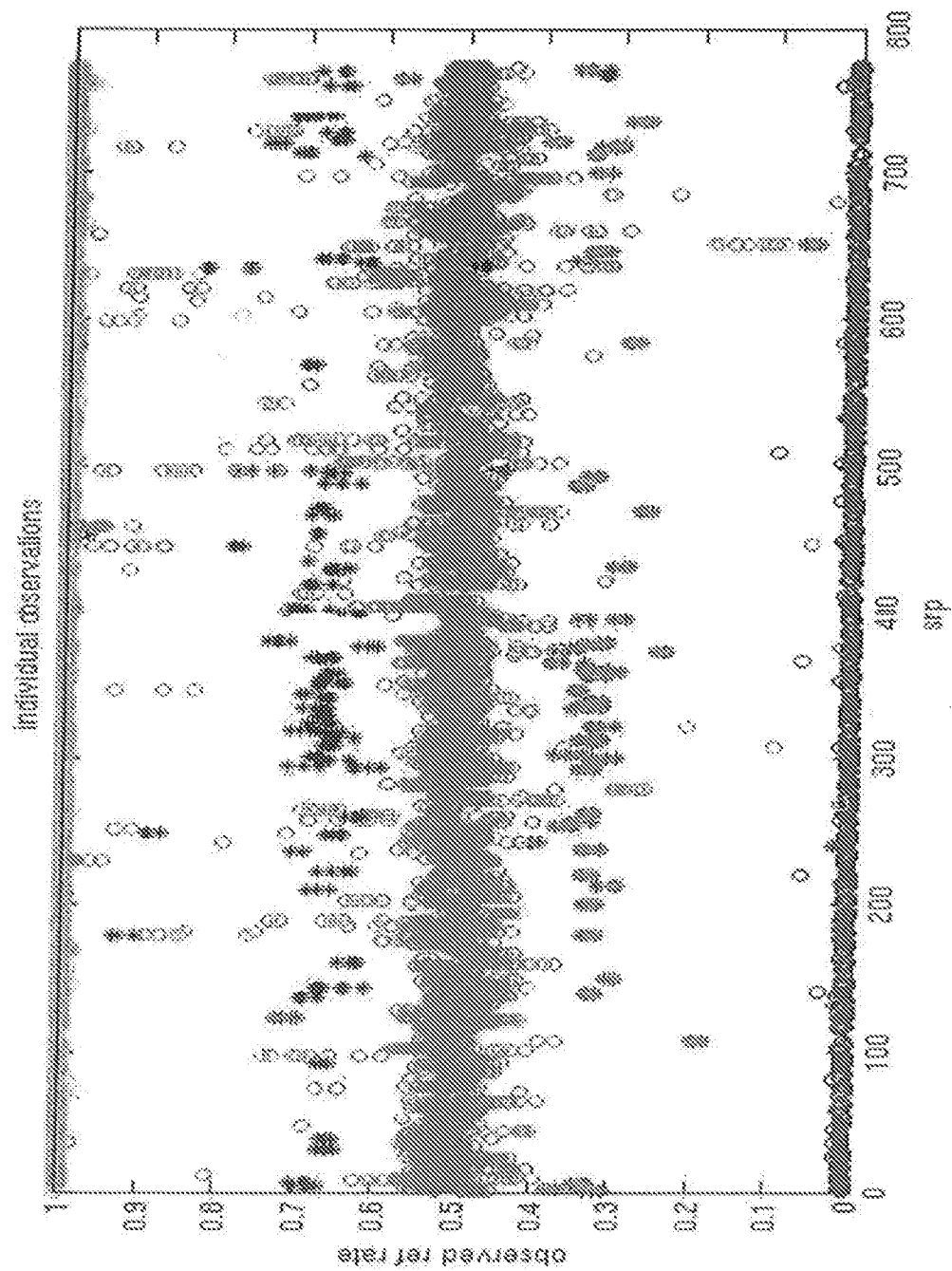


FIG 15

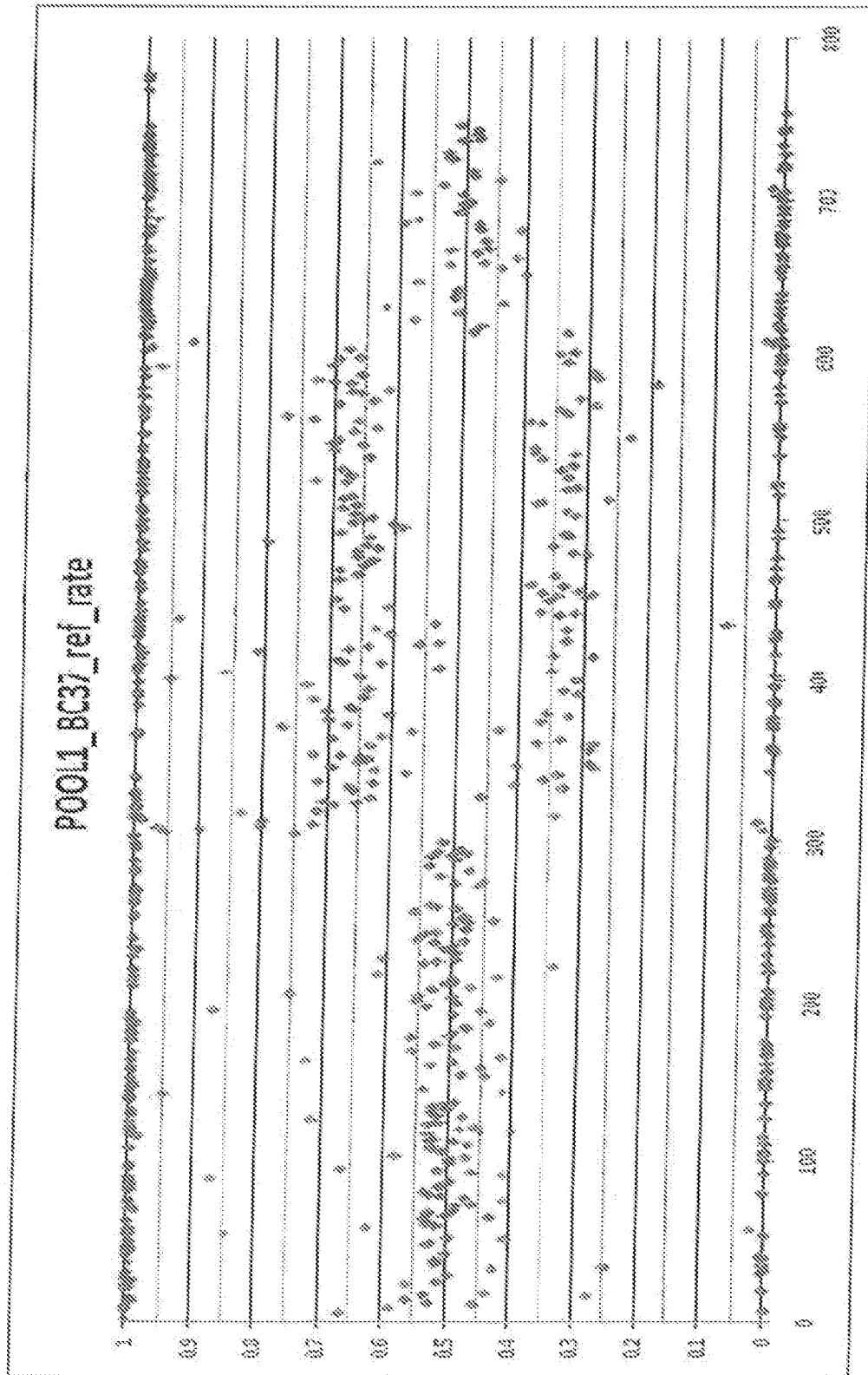


FIG.16

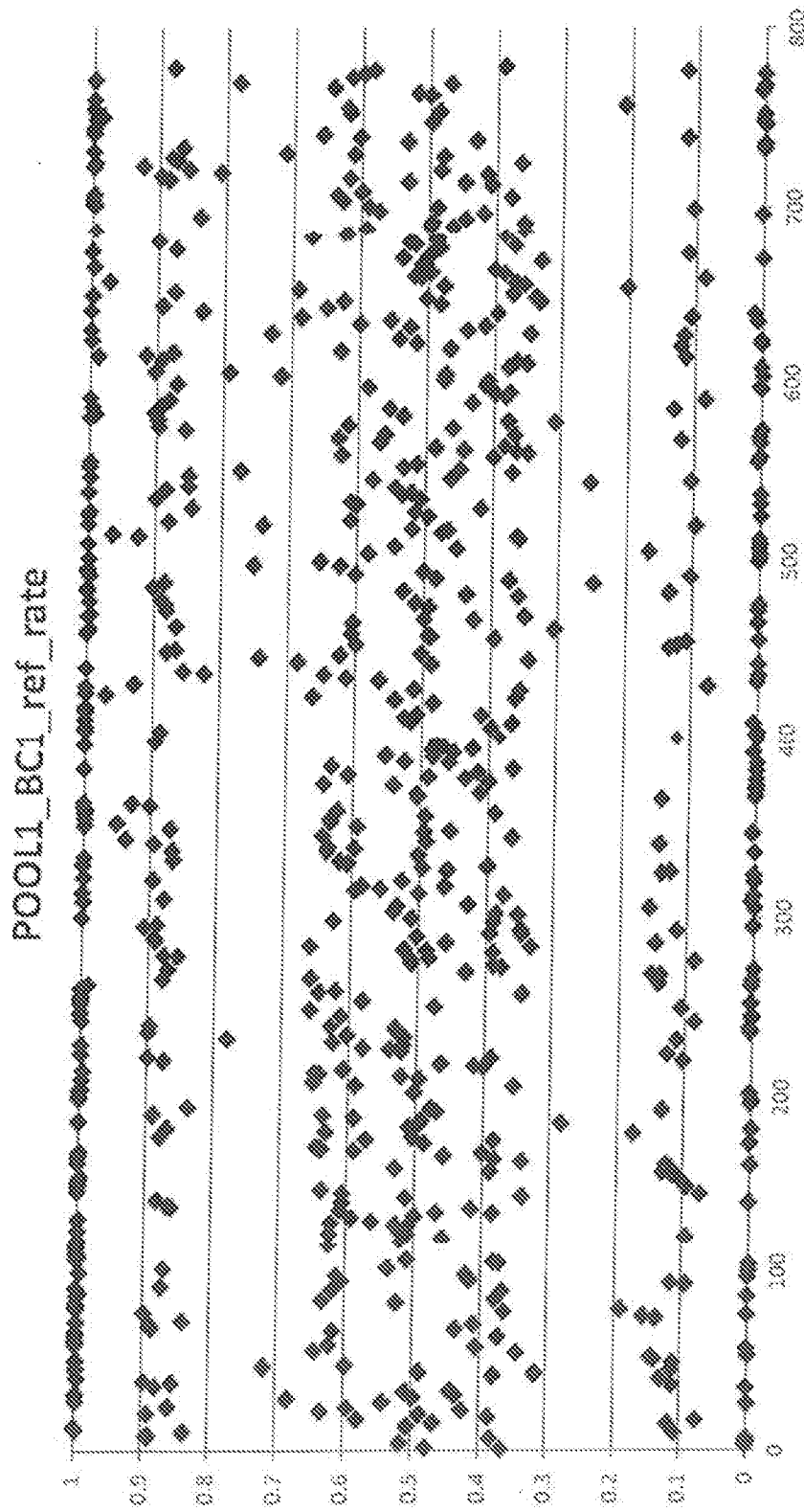


FIG. 17A

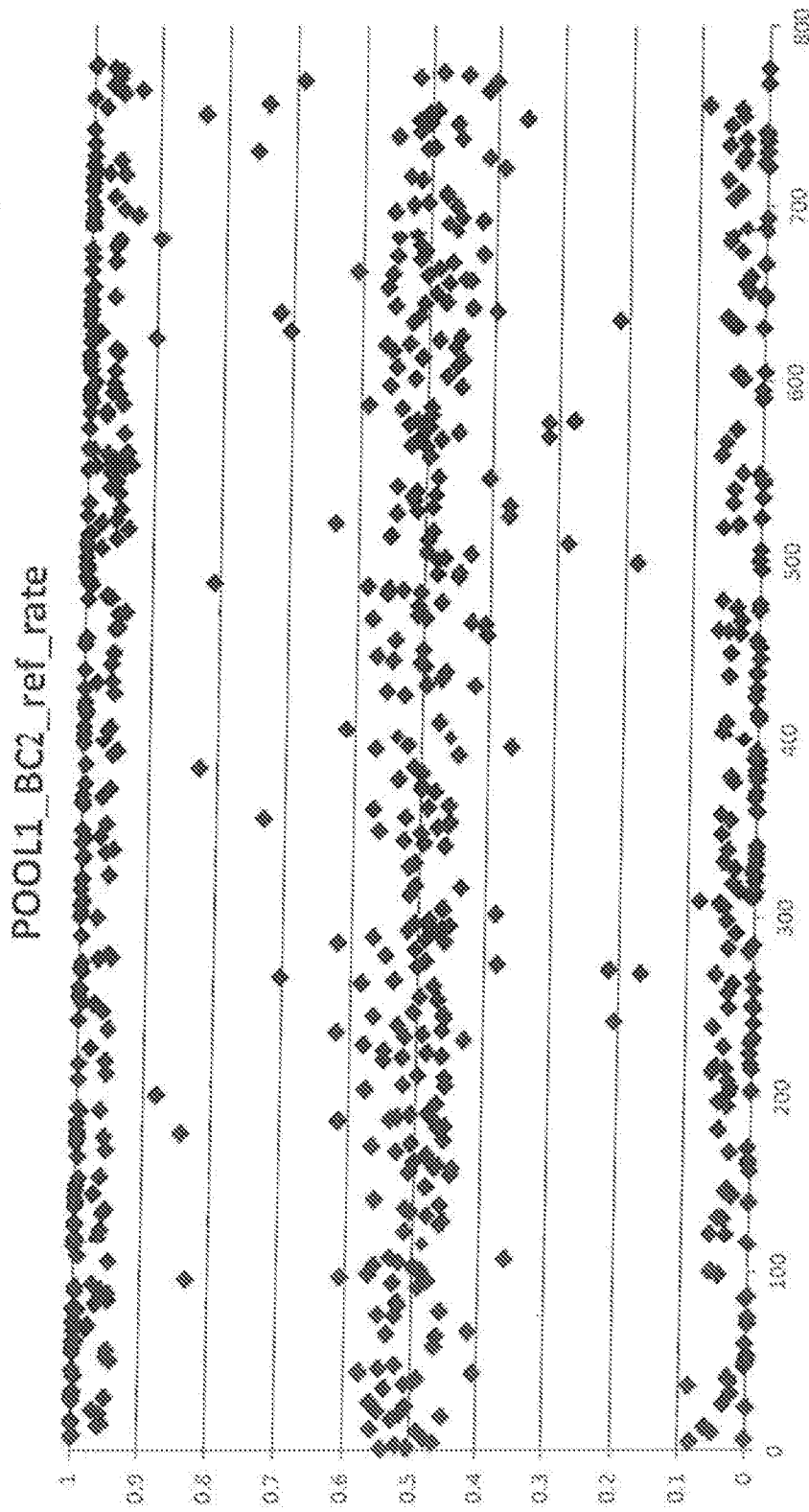


FIG. 17B

No fetal DNA

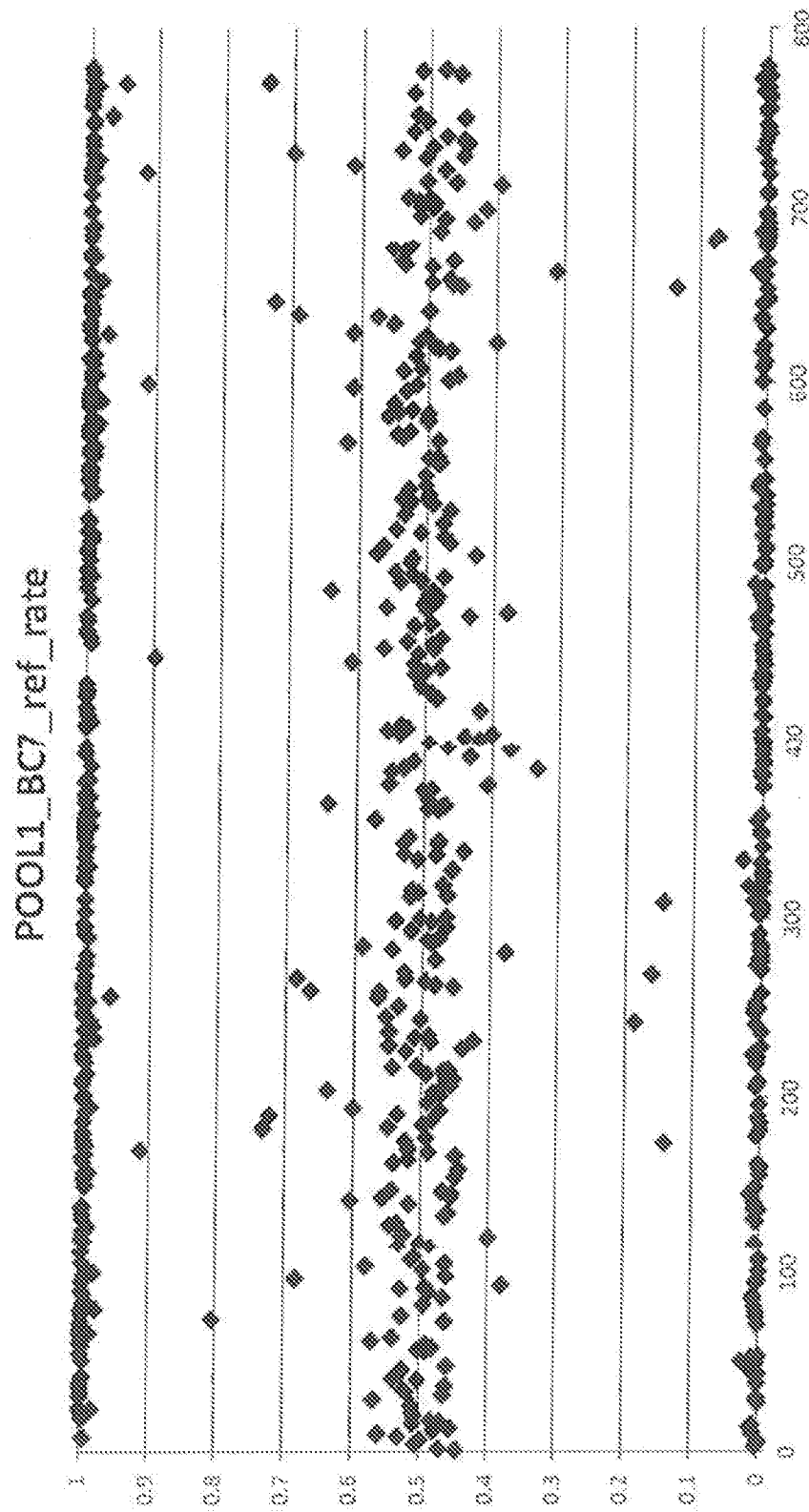


FIG. 17C

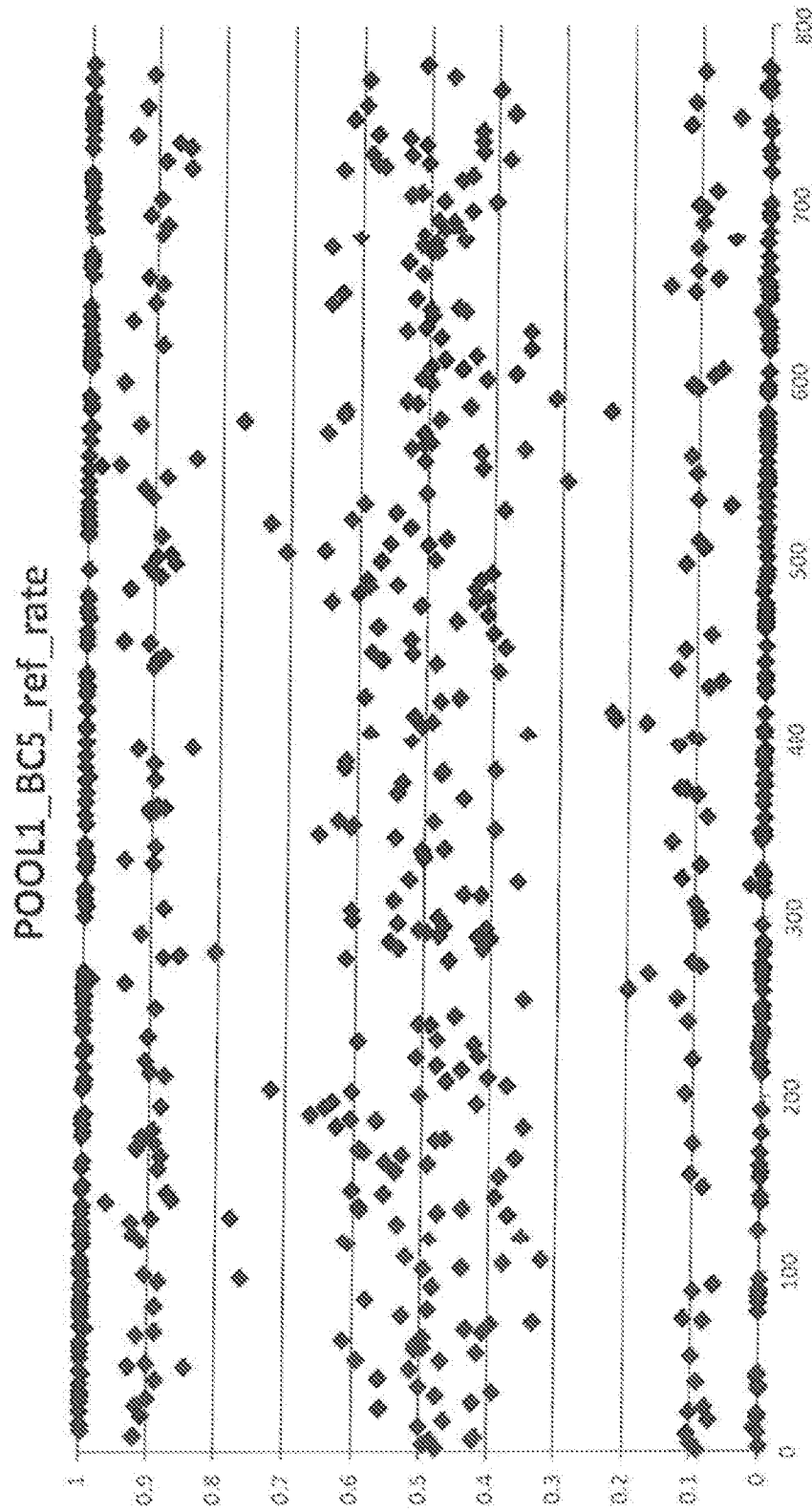


FIG. 17D

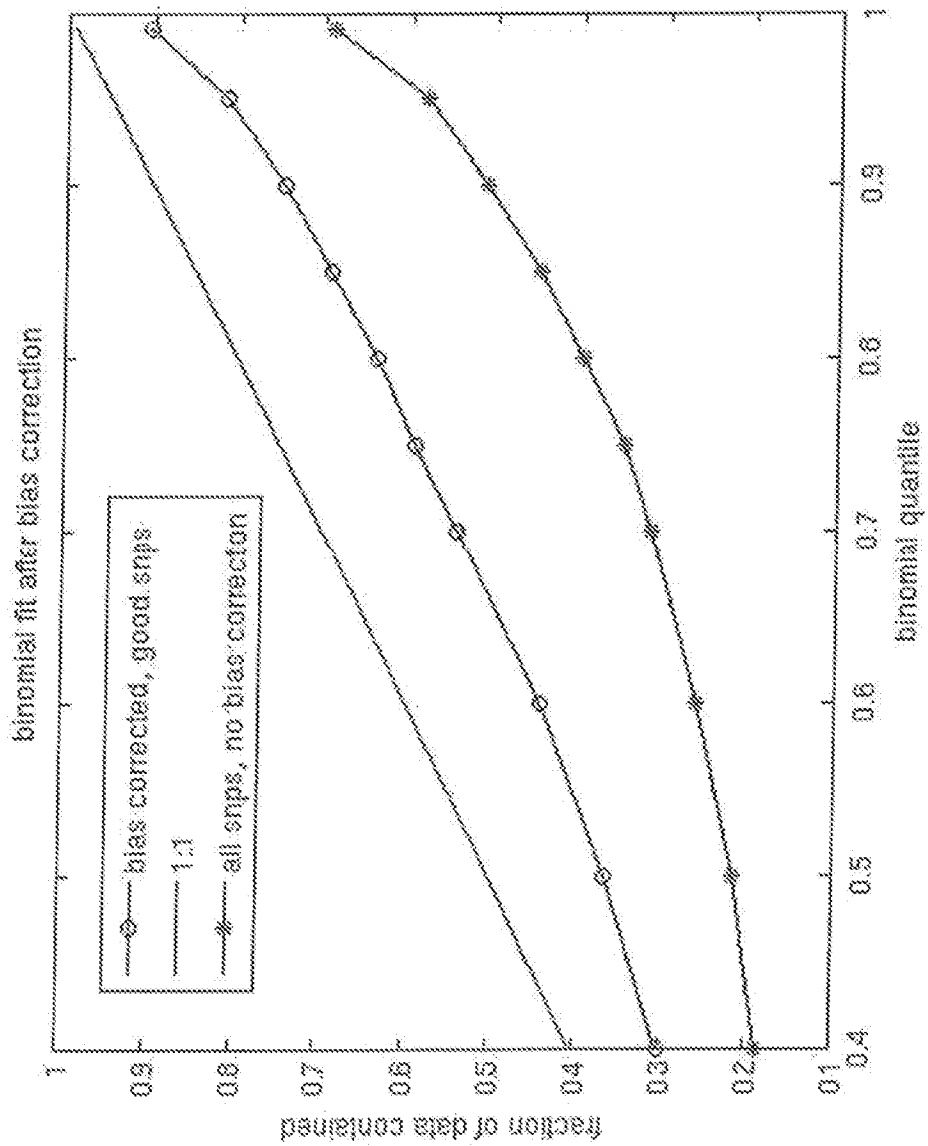


FIG. 18

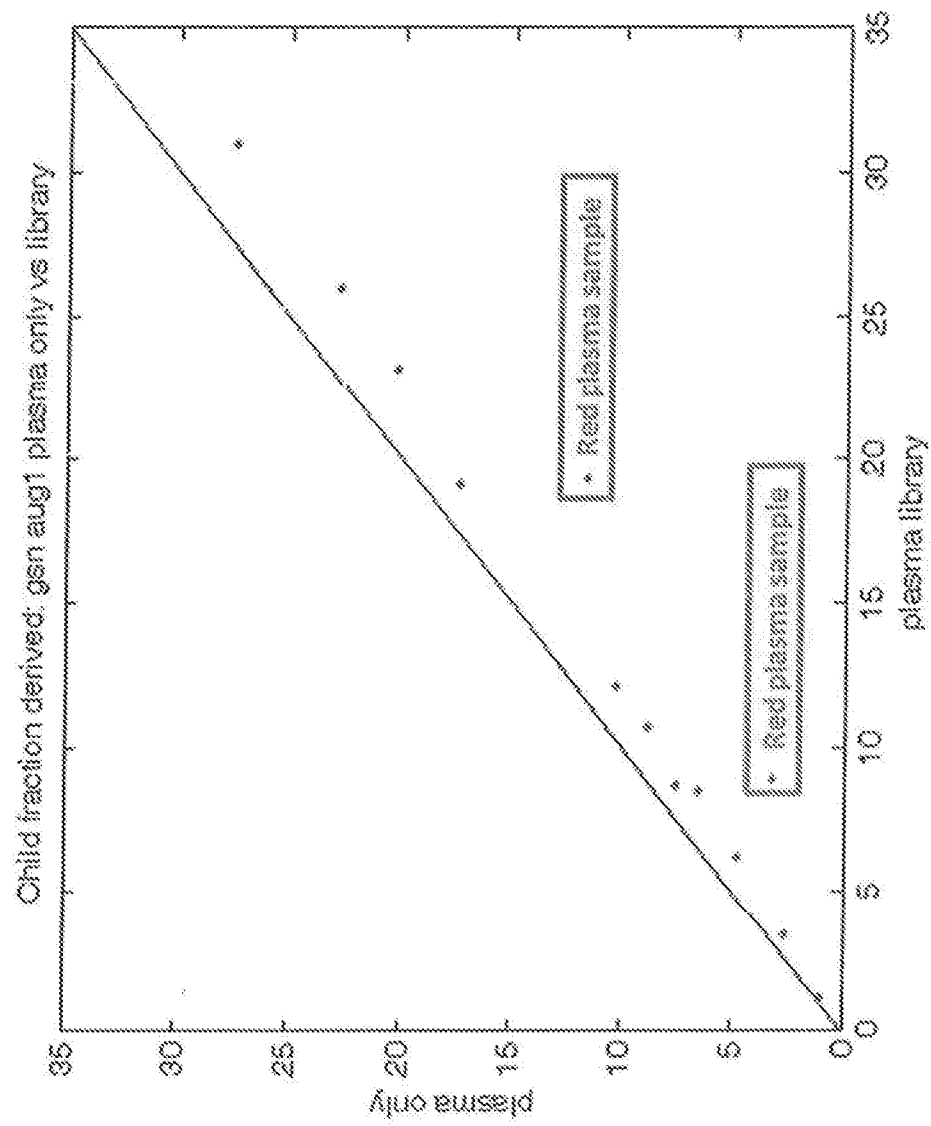


FIG. 19

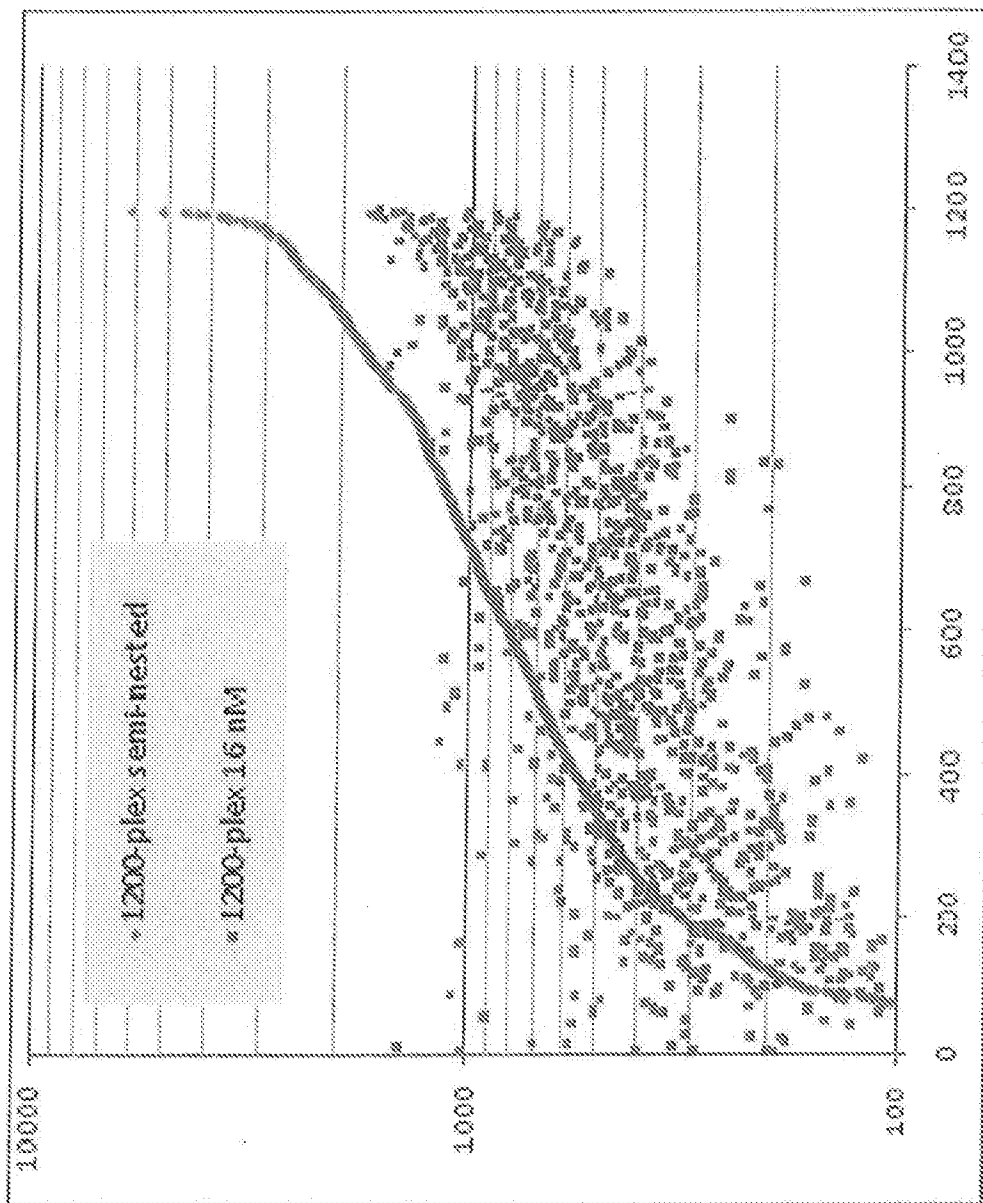


FIG. 20

DOR range direct 1200-p|ex, 10 nM, 10 min TA

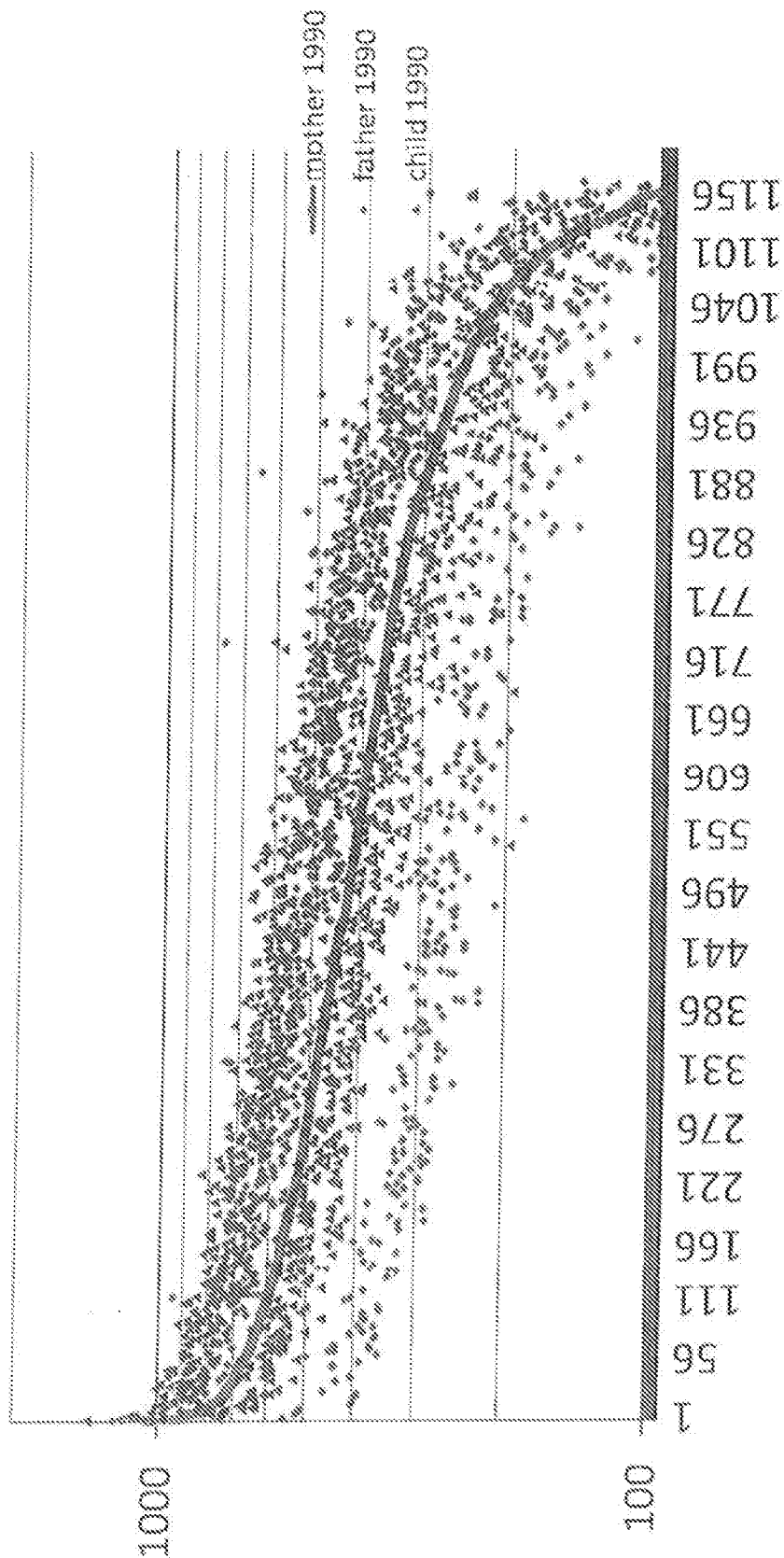


FIG. 2I

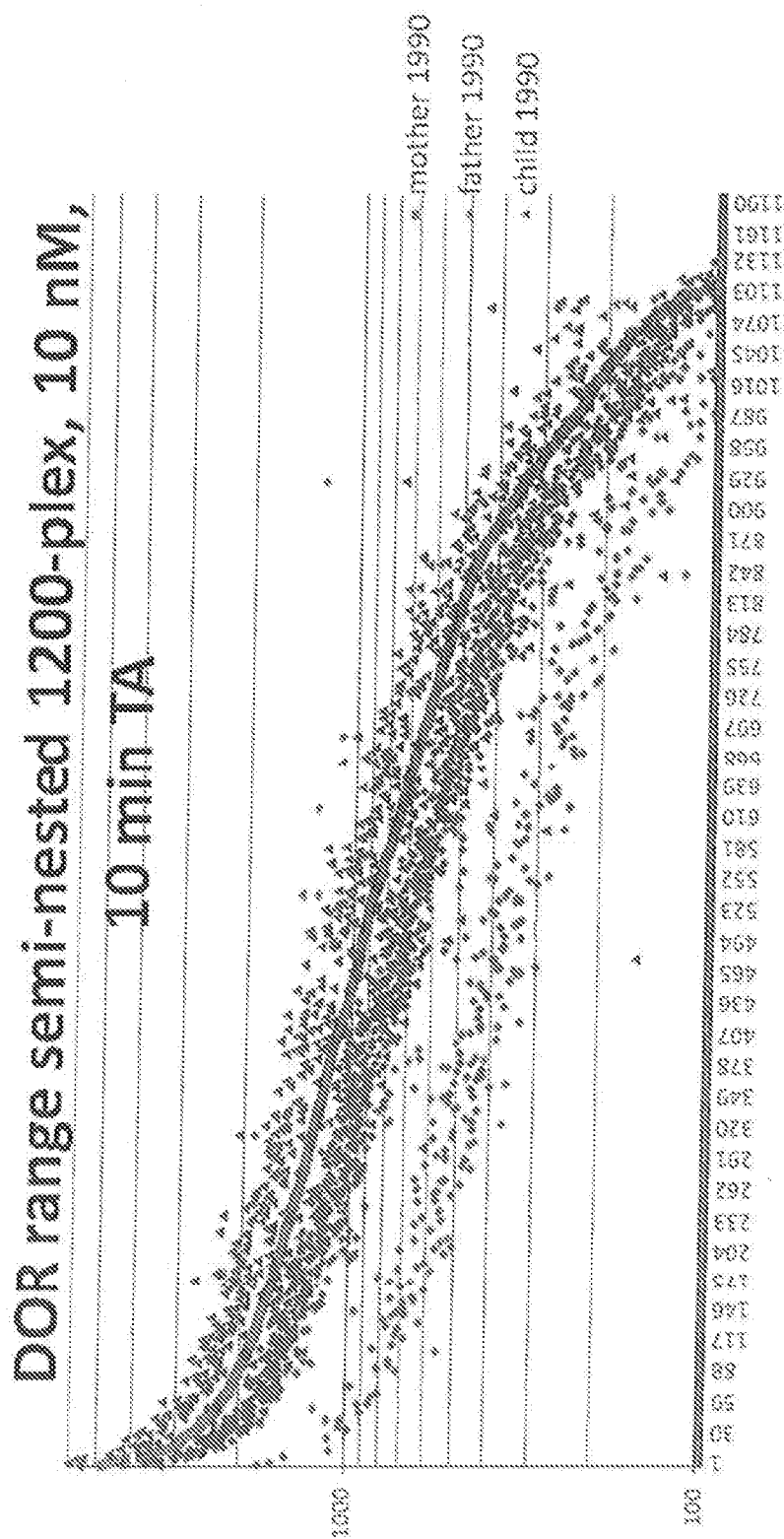


FIG. 22

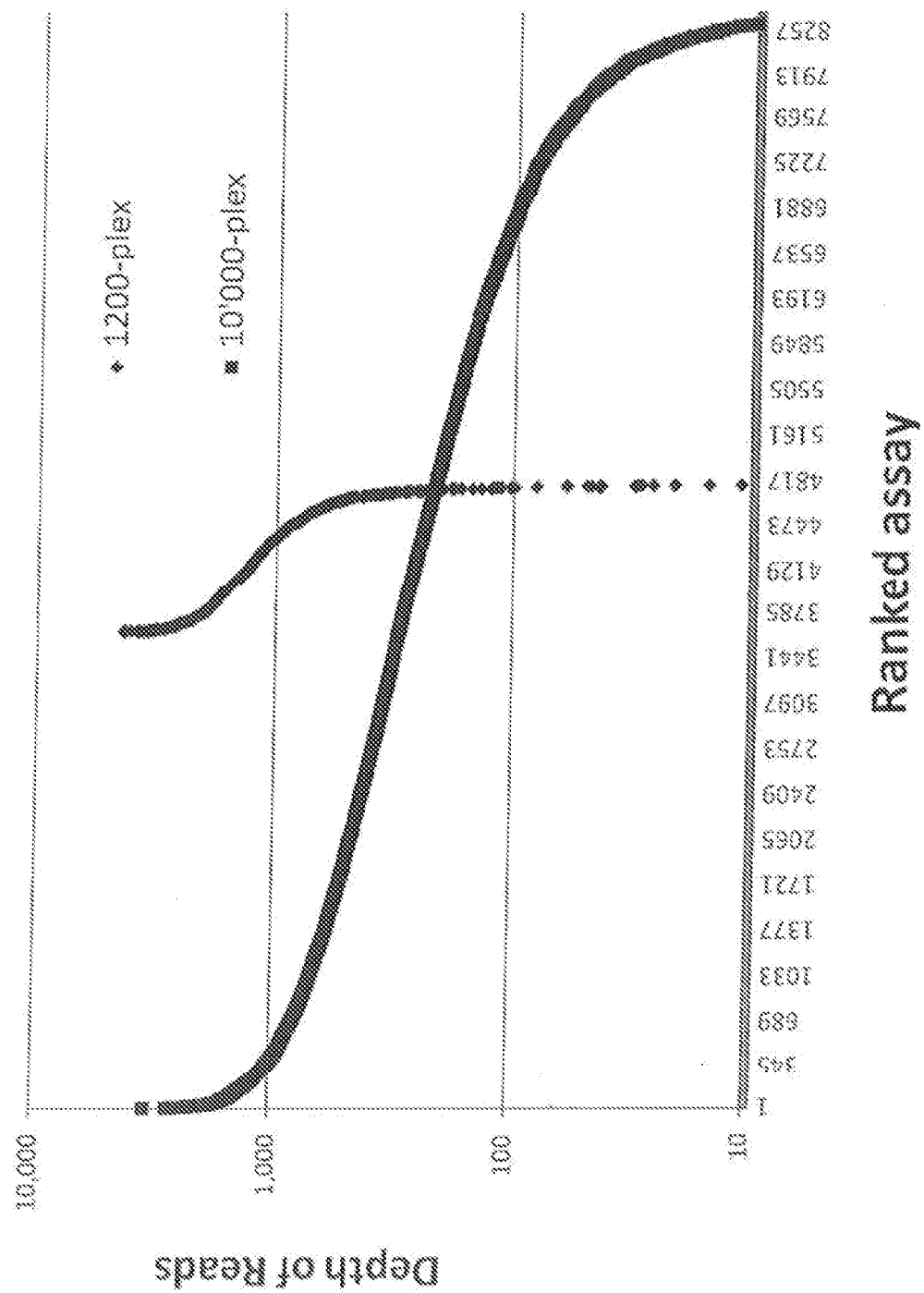


FIG. 23

DOR normalized vs DOR 46XY, 3cells

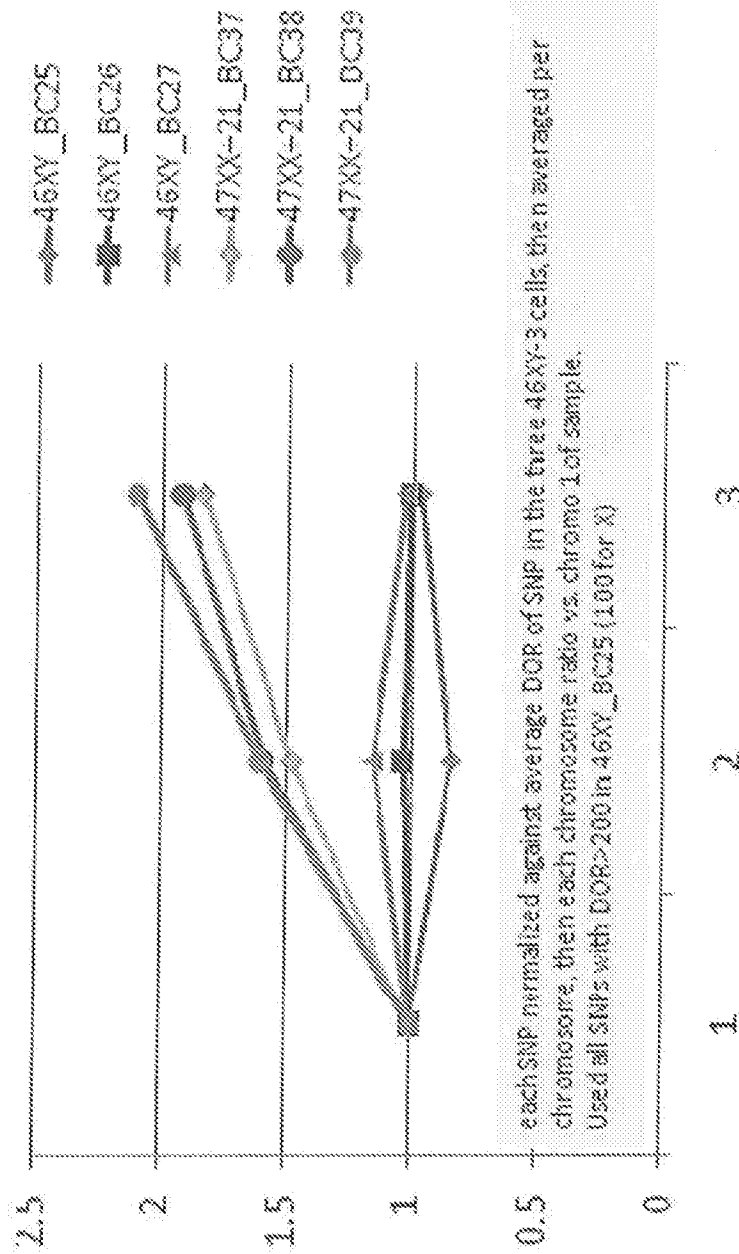


FIG. 24

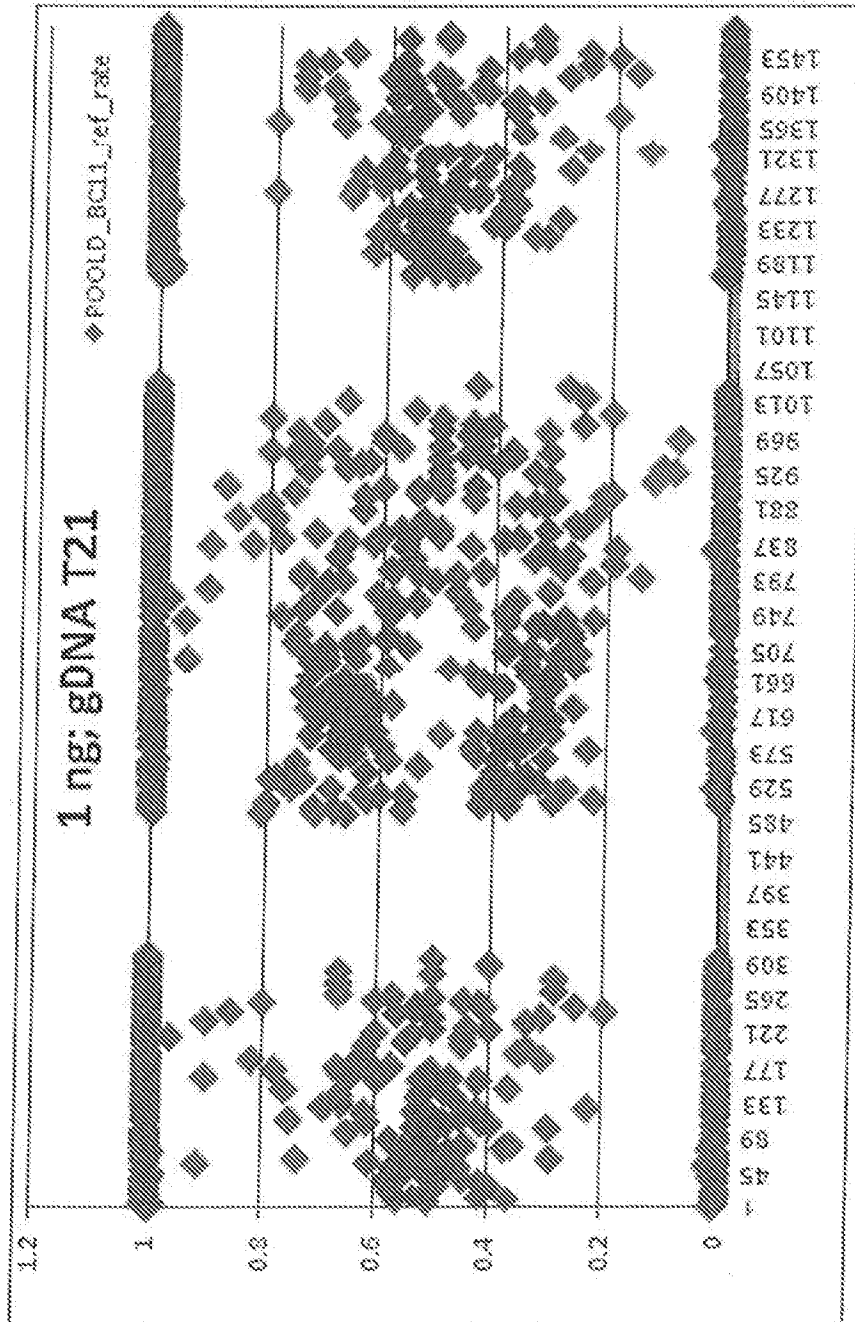


FIG. 25A

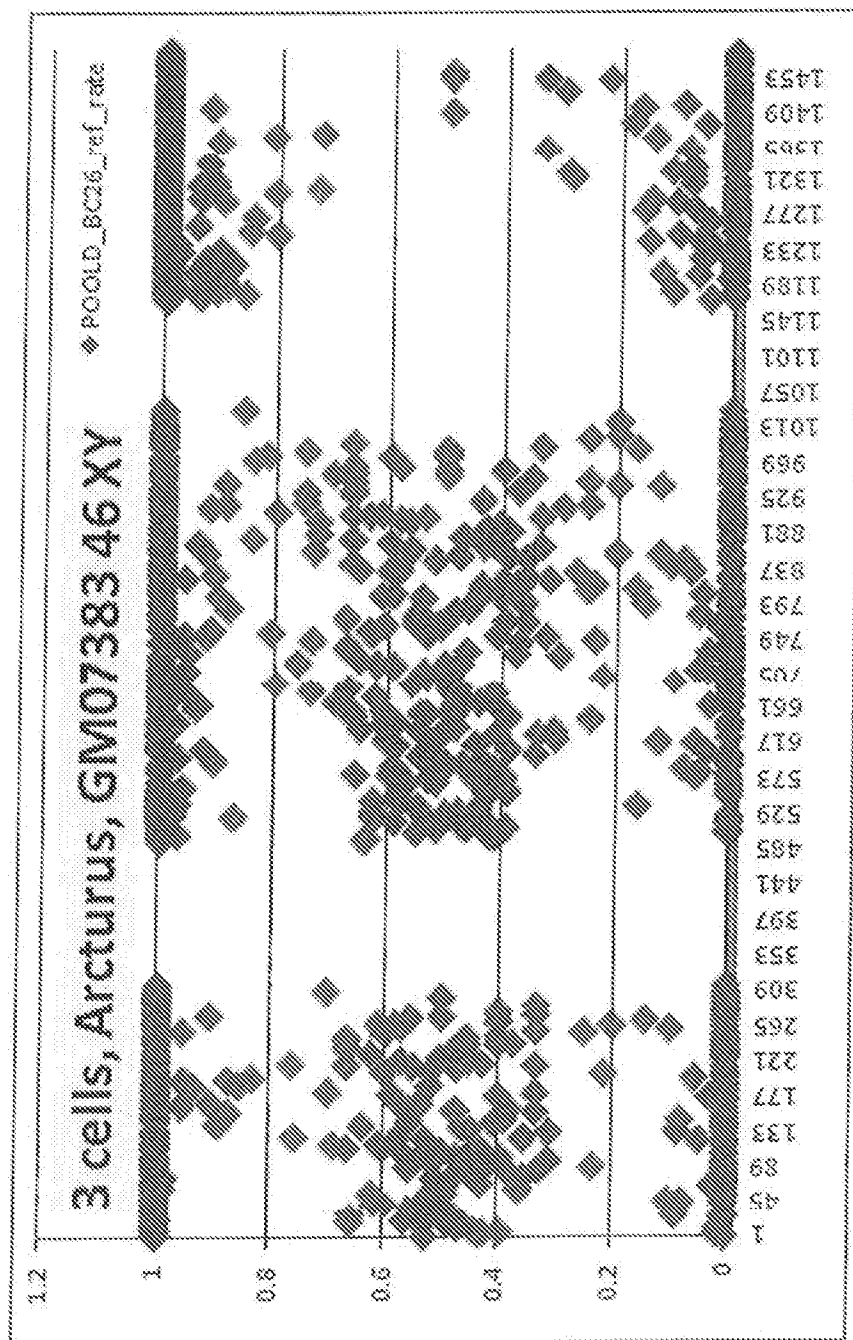


FIG. 25B

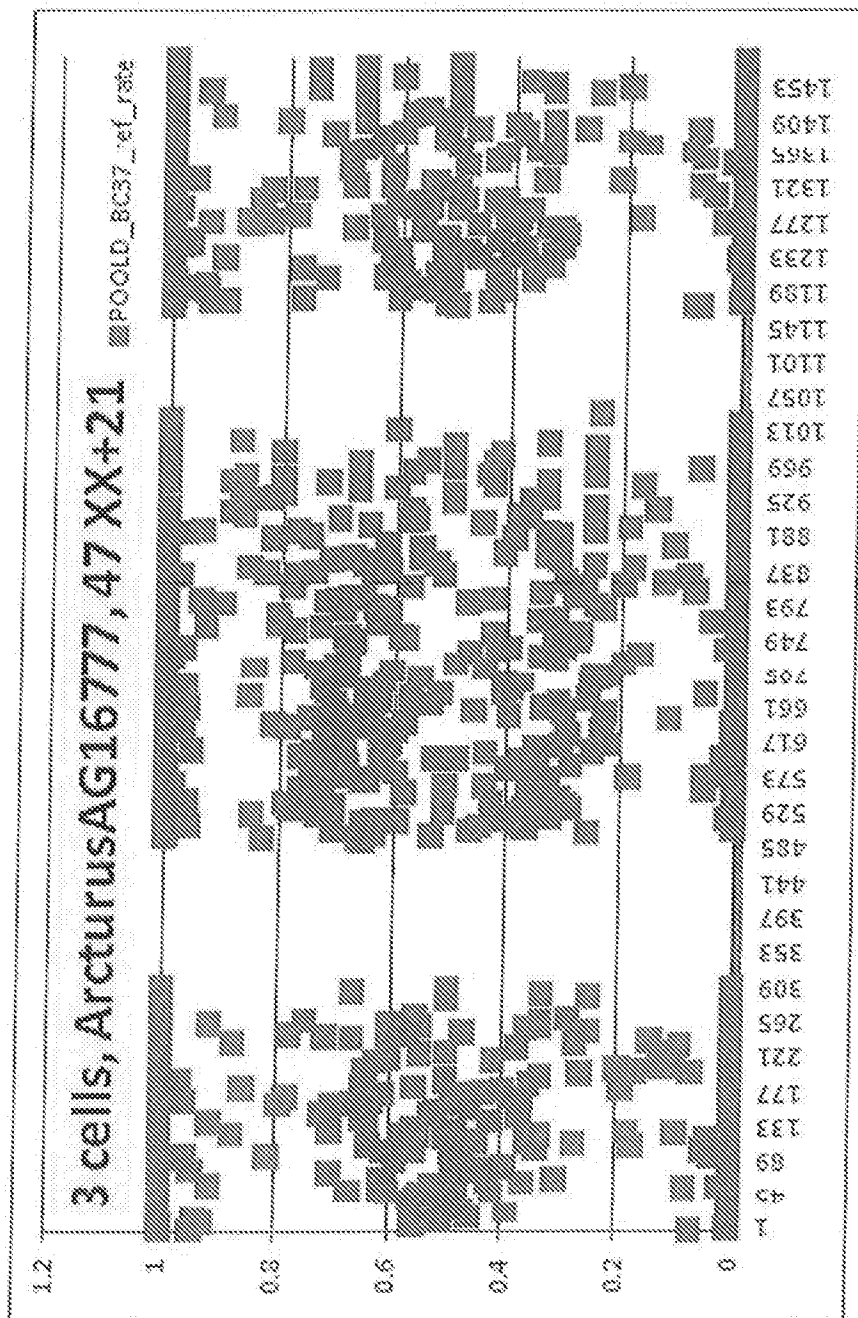


FIG. 25C

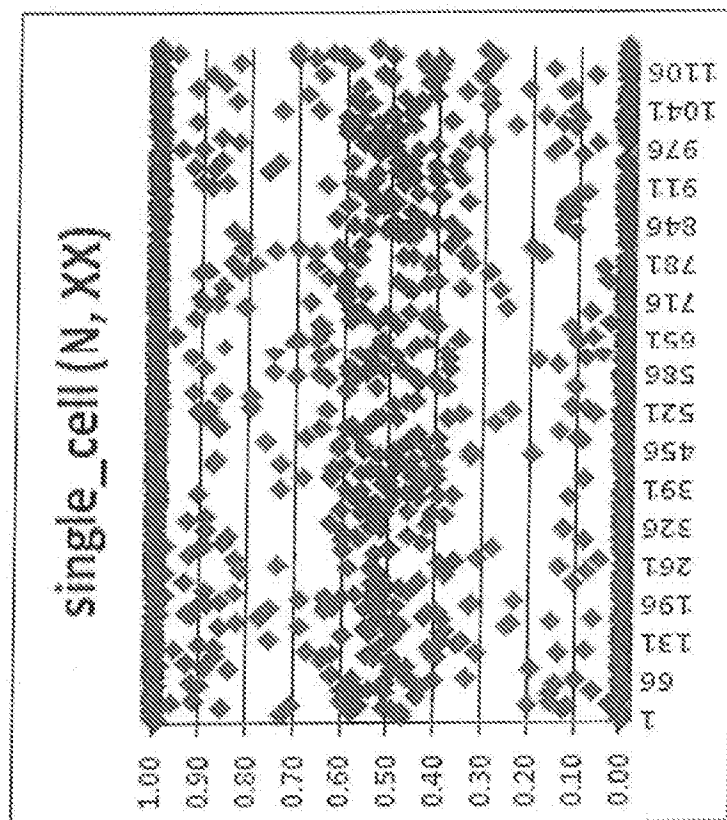


FIG. 26B

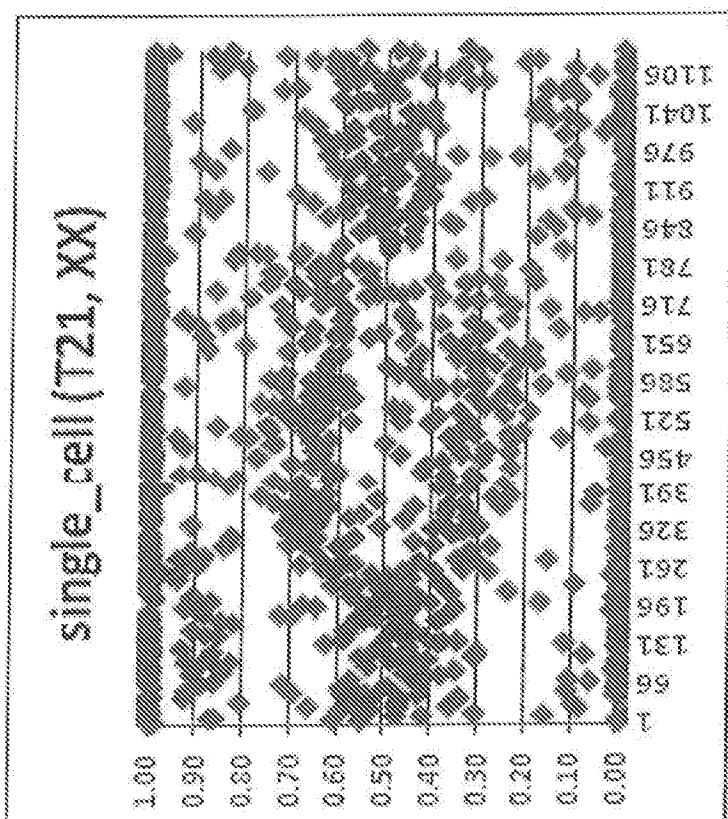


FIG. 26A

Minor Allele Frequencies for NPD Targets, Chromosome 21

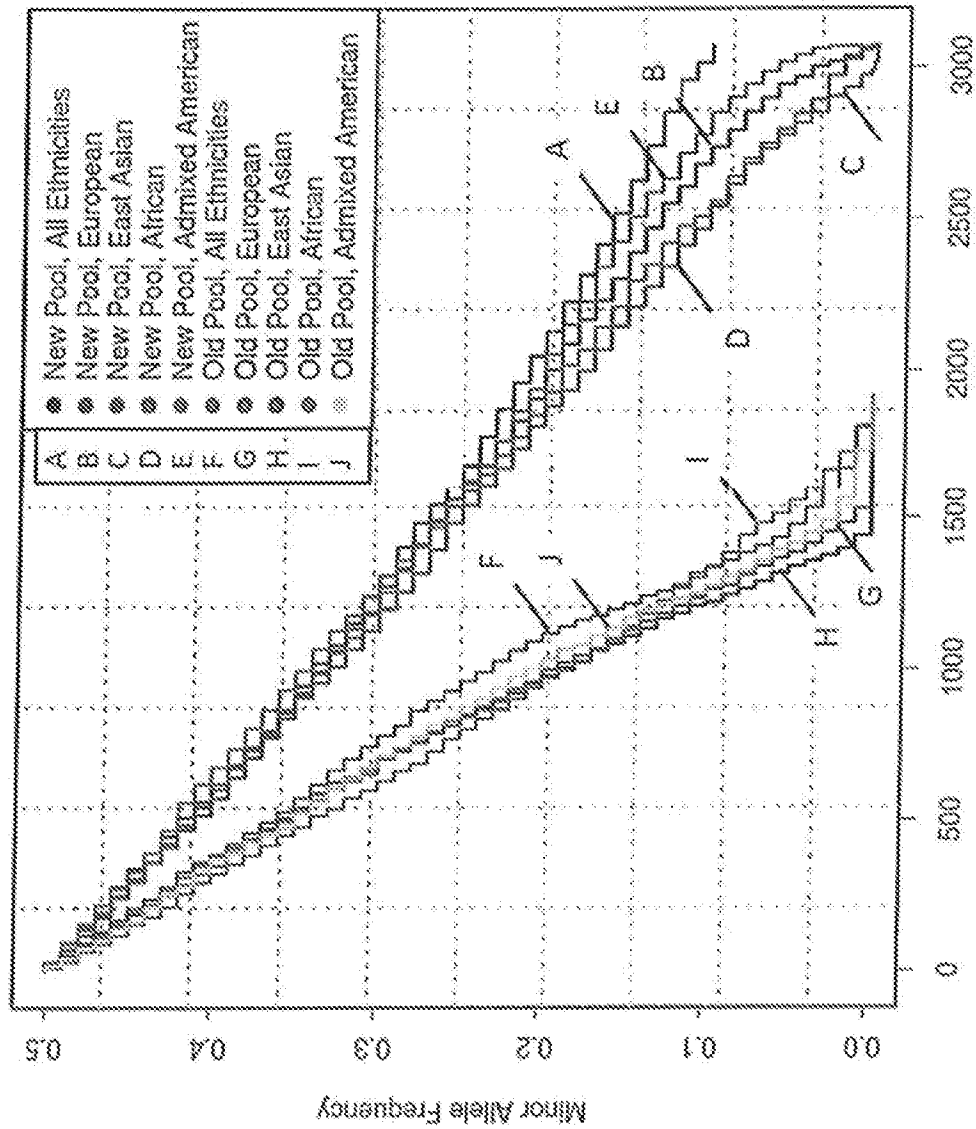


FIG. 27

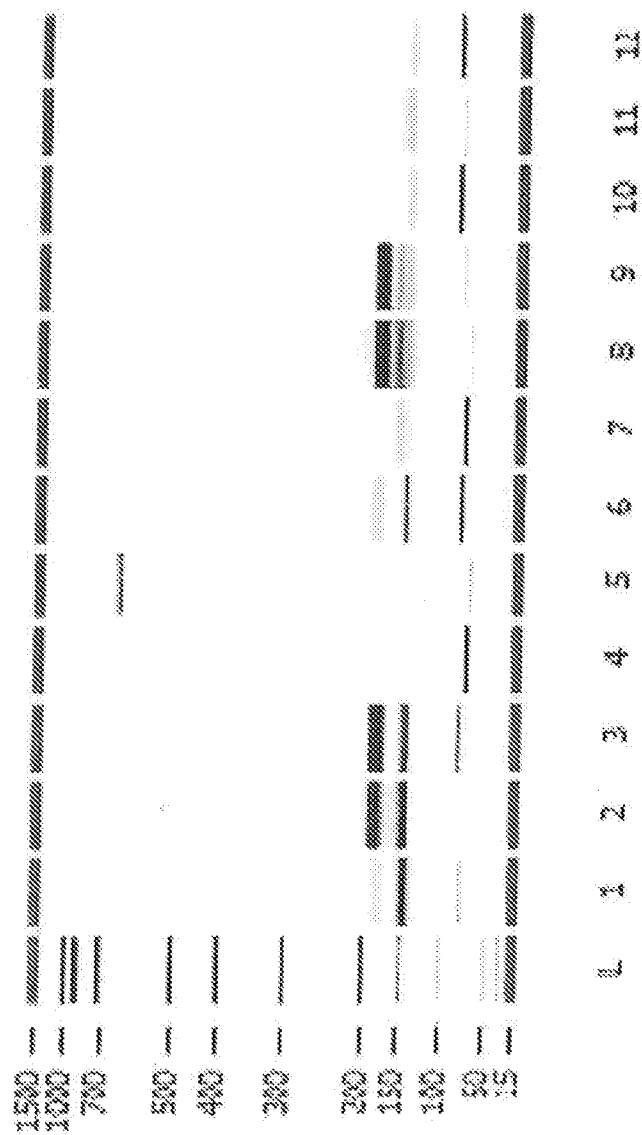


FIG. 28A

Direct PCR 2.4k 24ng 10nM

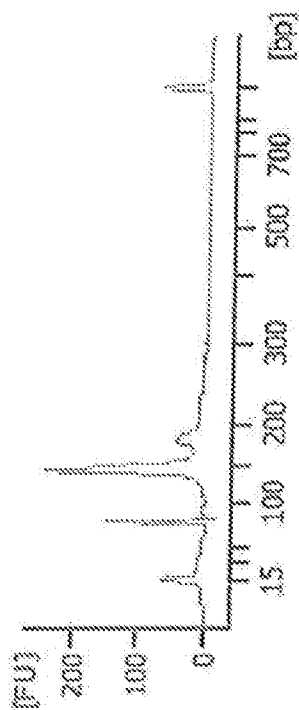


FIG. 28B

Direct PCR 2.4k 80ng 10nM

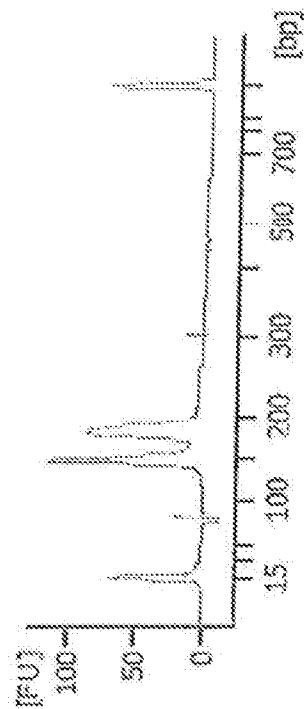


FIG. 28C

Direct PCR 2.4k 250ng 10nM

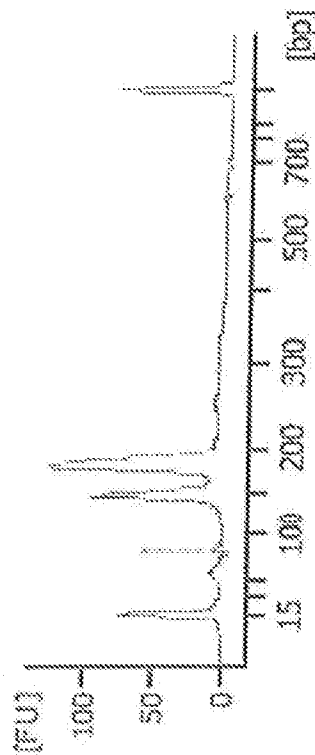


FIG. 28D

Direct PCR 2.4k 24ng 1nM

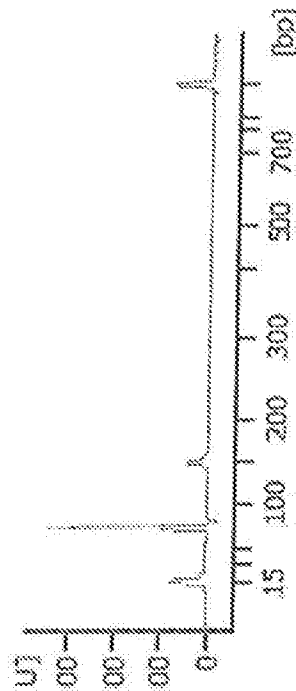


FIG. 28E

Direct PCR 2.4k 80ng 1nM



FIG. 28F

Direct PCR 2.4k 250ng 1nM

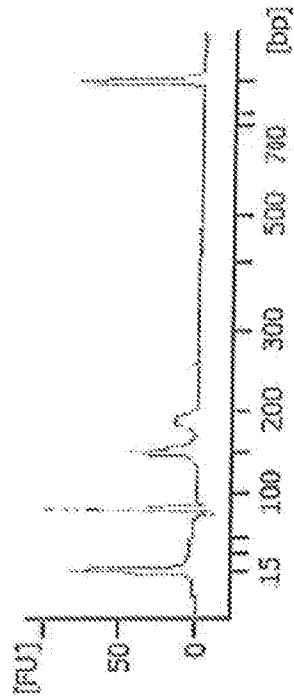


FIG. 28G

Direct PCR 20k 24ng 10nM

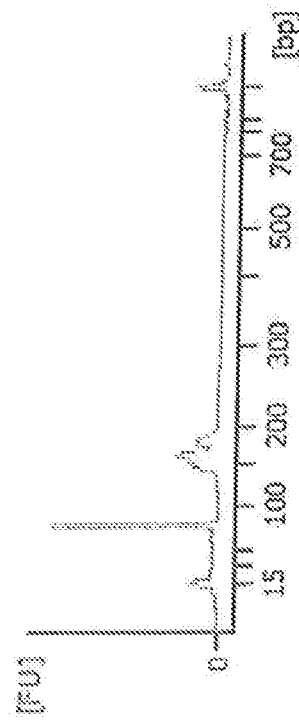


FIG. 28H

Direct PCR 20k 80ng 10nM

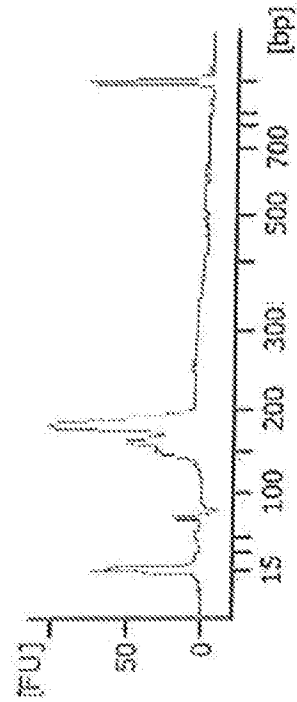


FIG. 28I

Direct PCR 20k 250ng 10nM

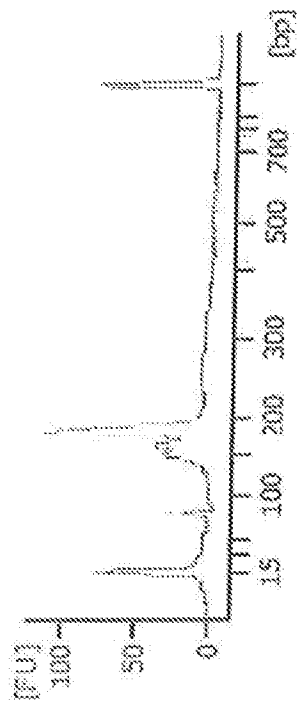


FIG. 28J

Direct PCR 20k 24ng 1nM

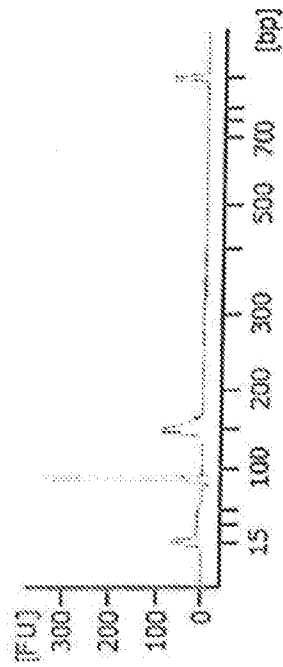


FIG. 28K

Direct PCR 20k 80ng 1nM

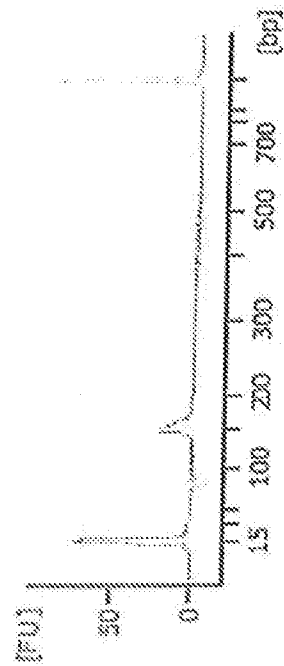


FIG. 28L

Direct PCR 20k 250ng 1nM

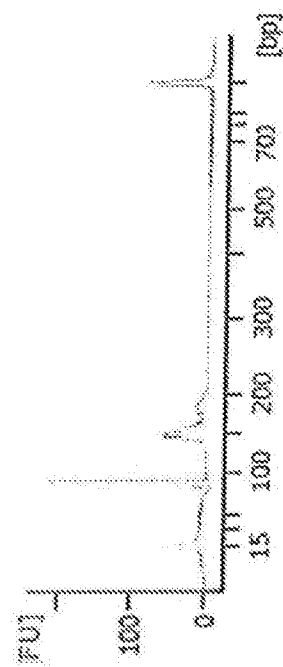
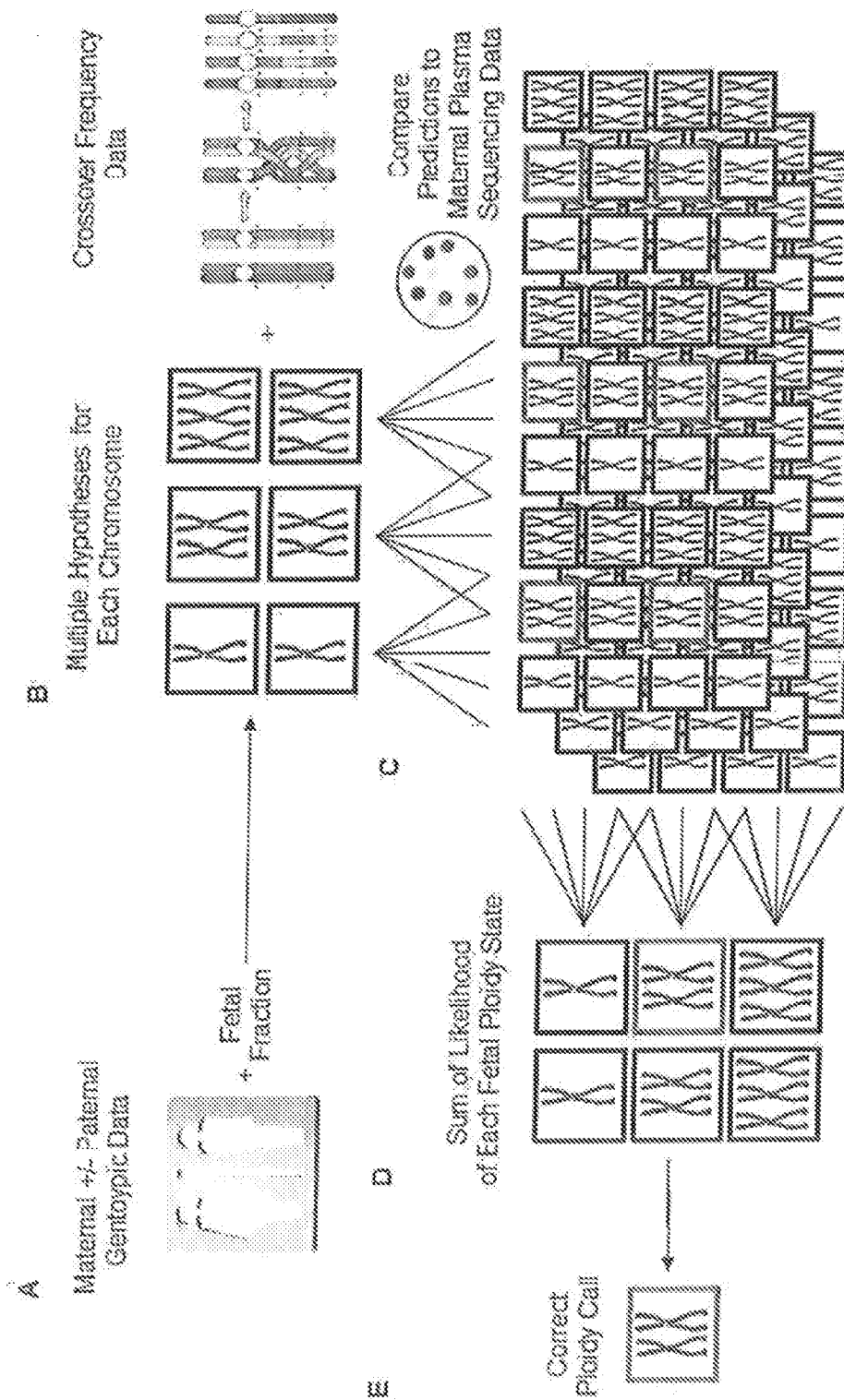
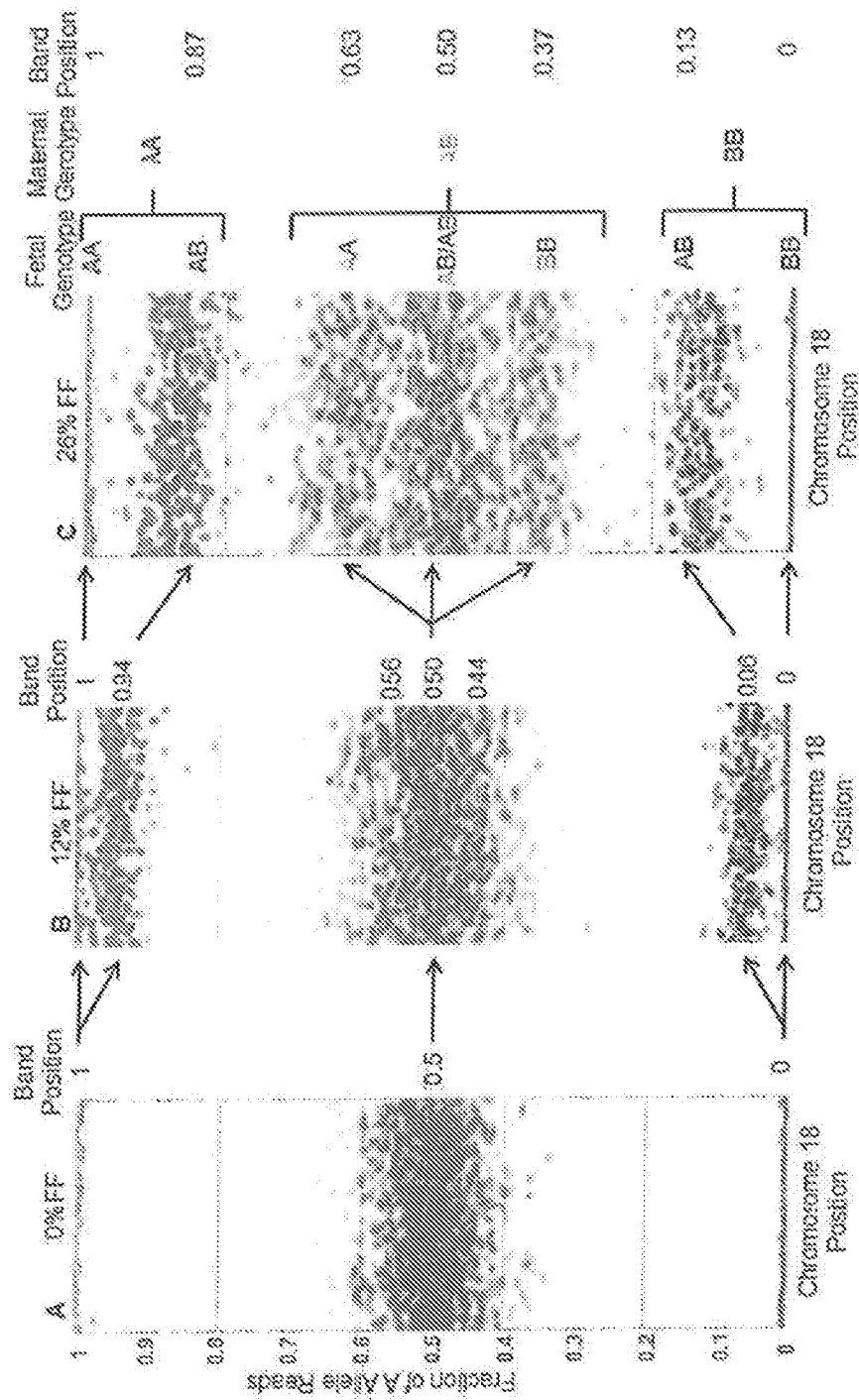


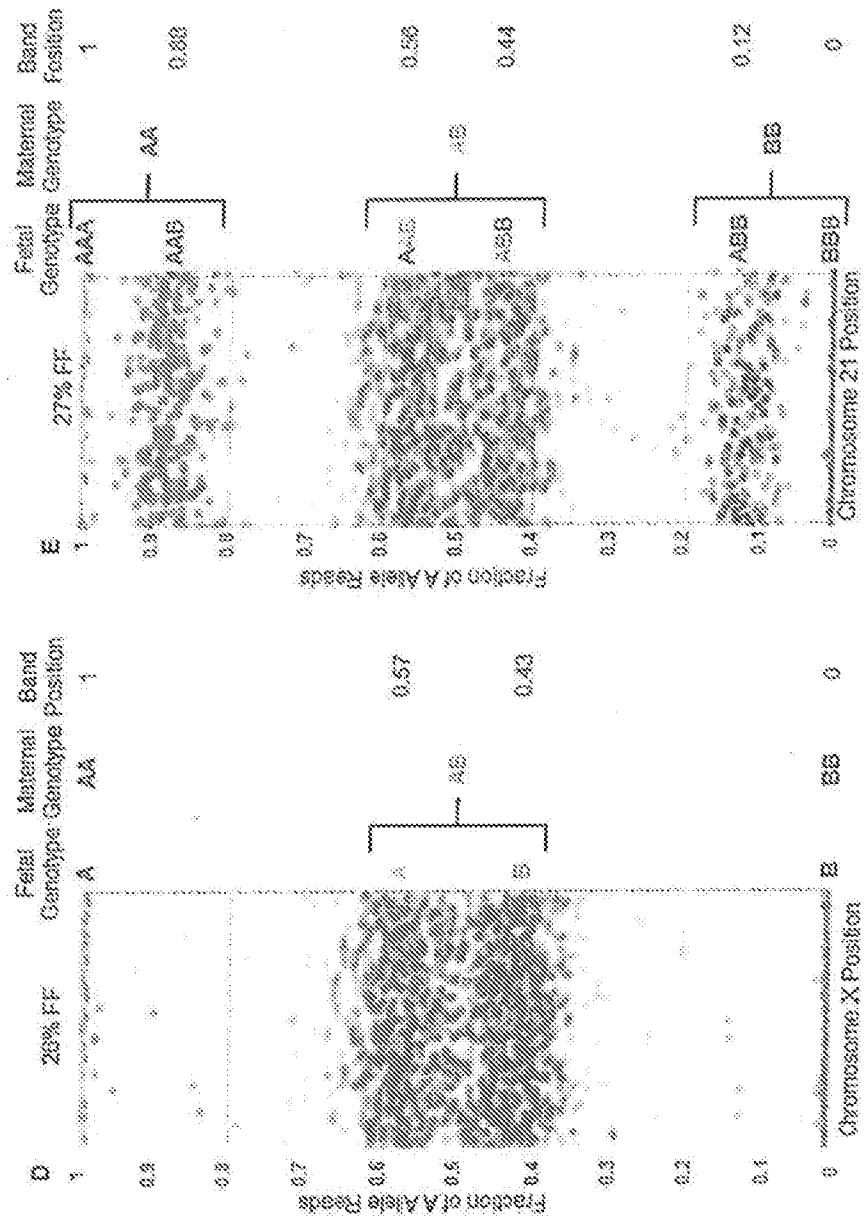
FIG. 28M



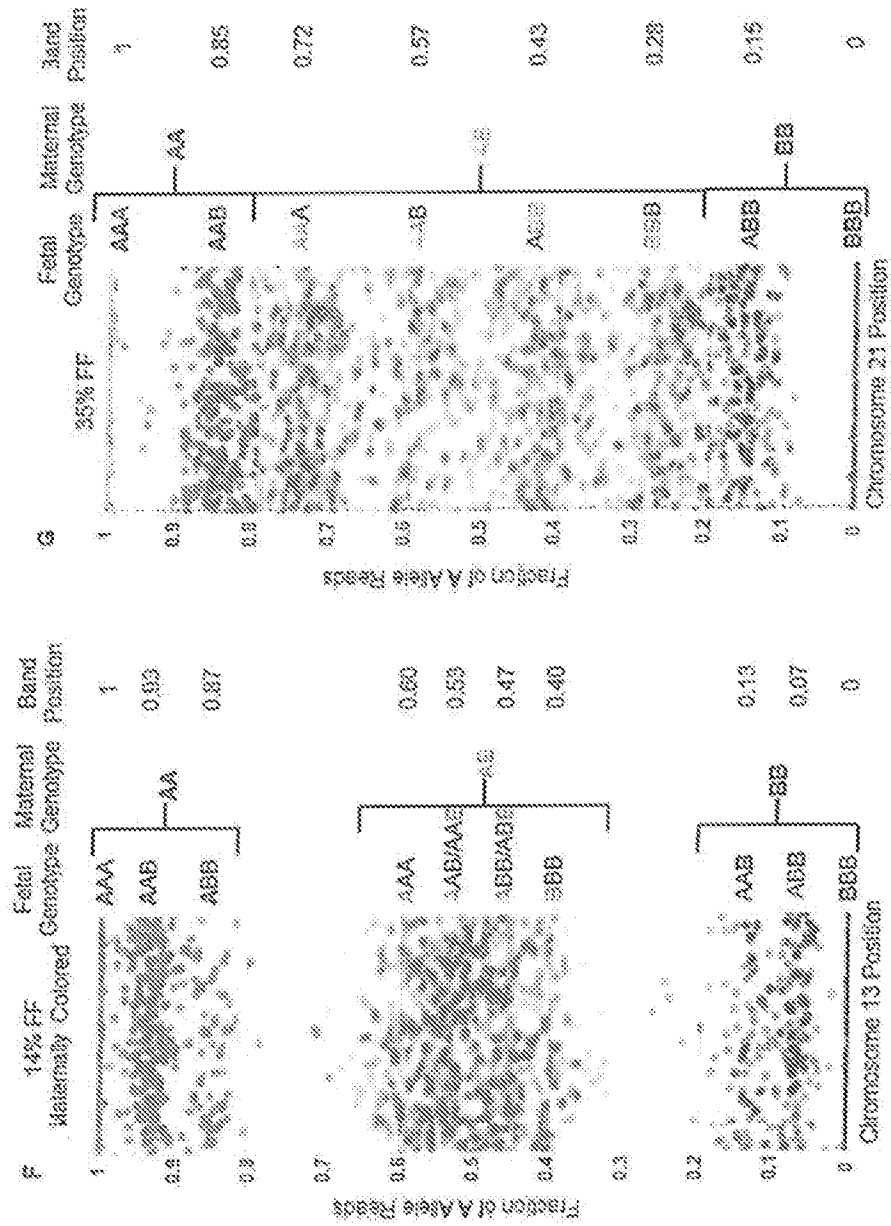
FIGS. 29A-29E



FIGS. 30A-30C



FIGS. 30D & 30E



FIGS. 30F & 30G

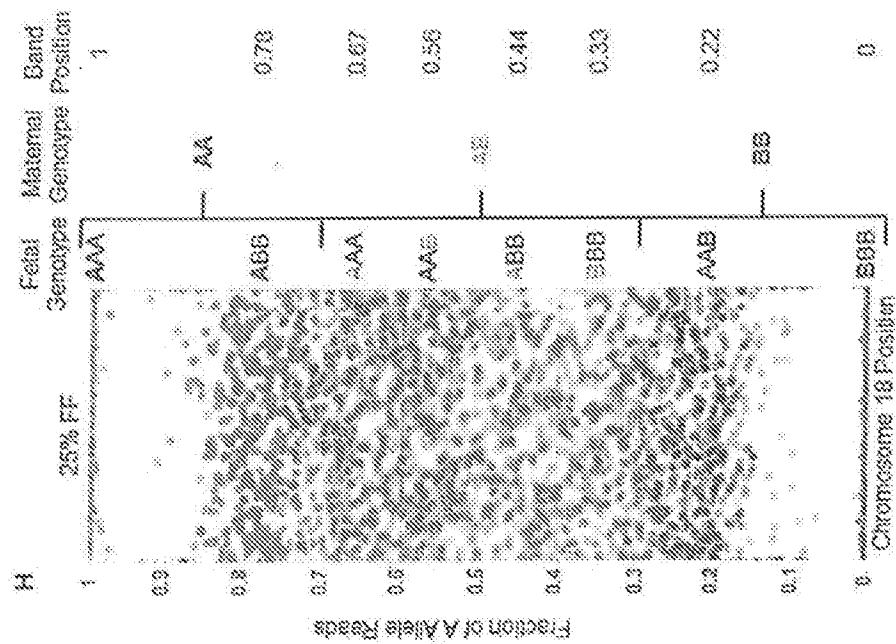
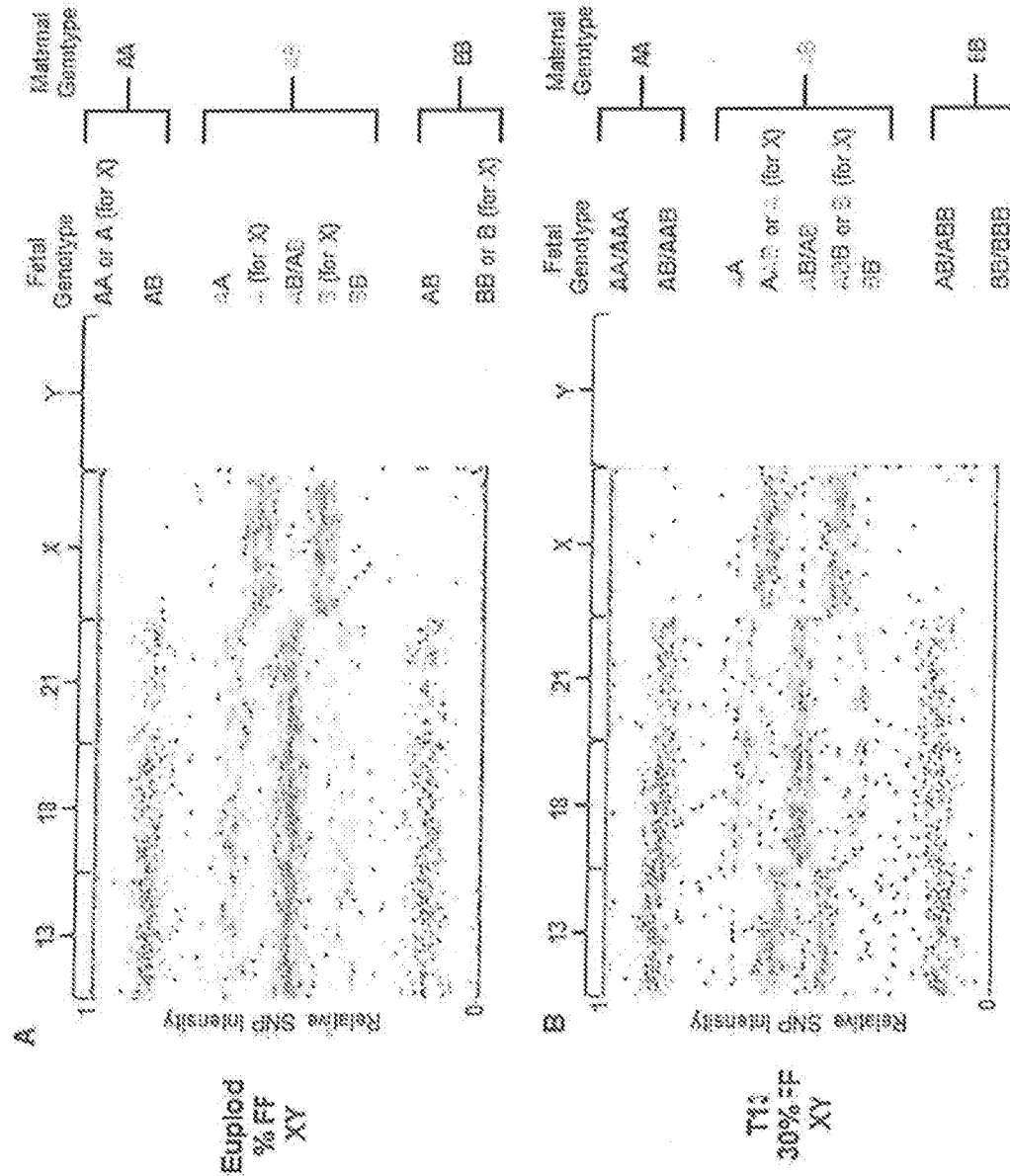
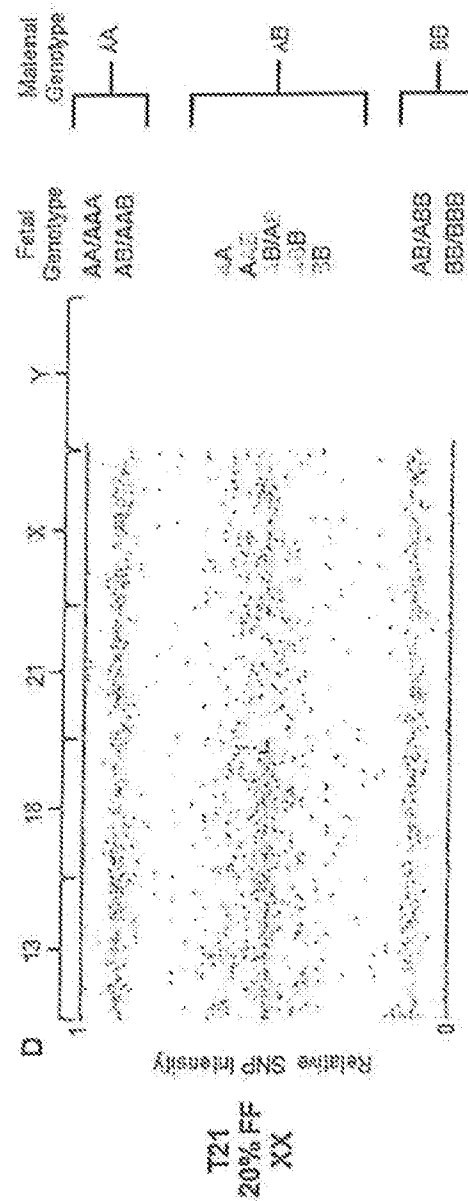
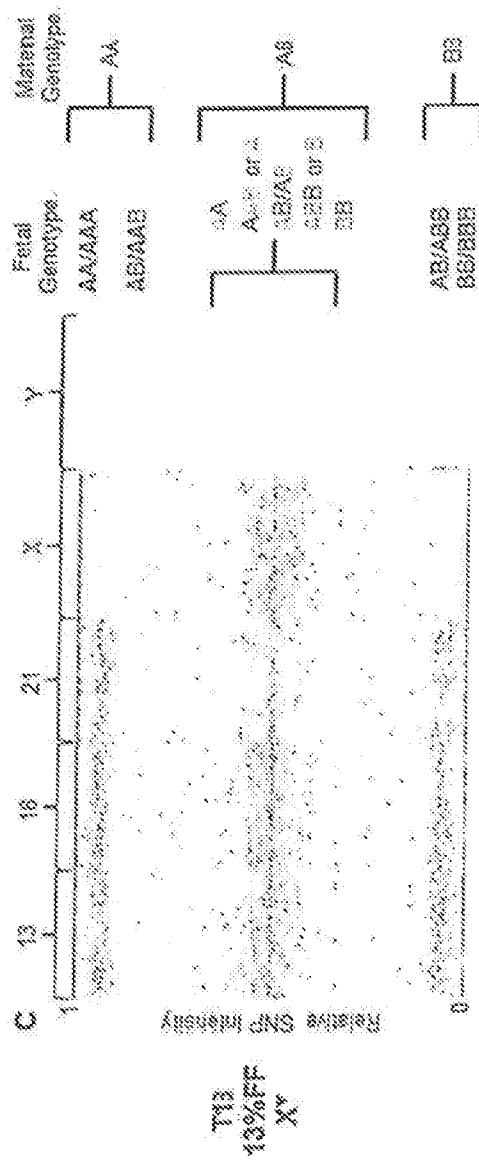


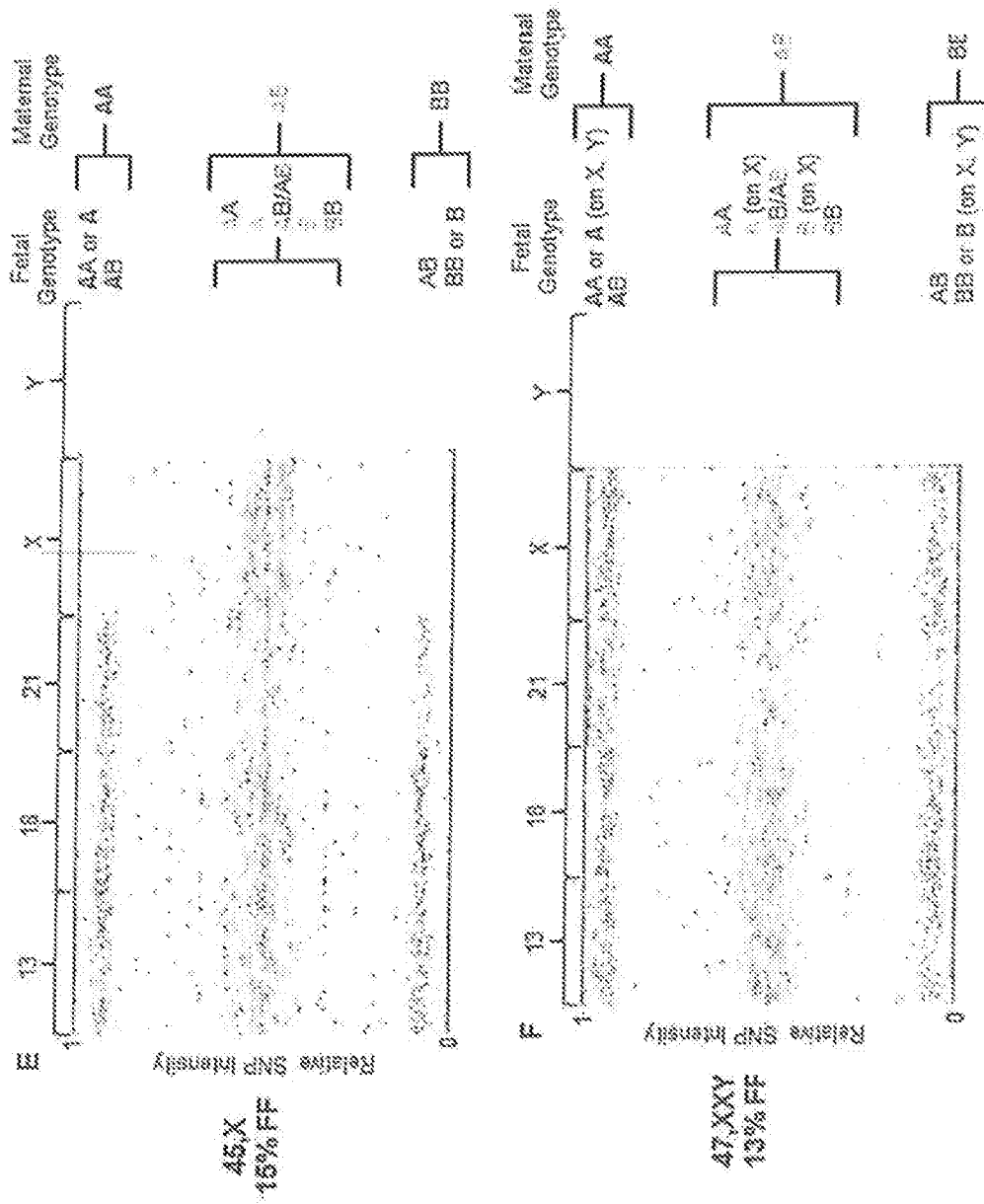
FIG. 30H



FIGS. 31A & 31B



FIGS. 31C & 31D



FIGS. 31E & 31F

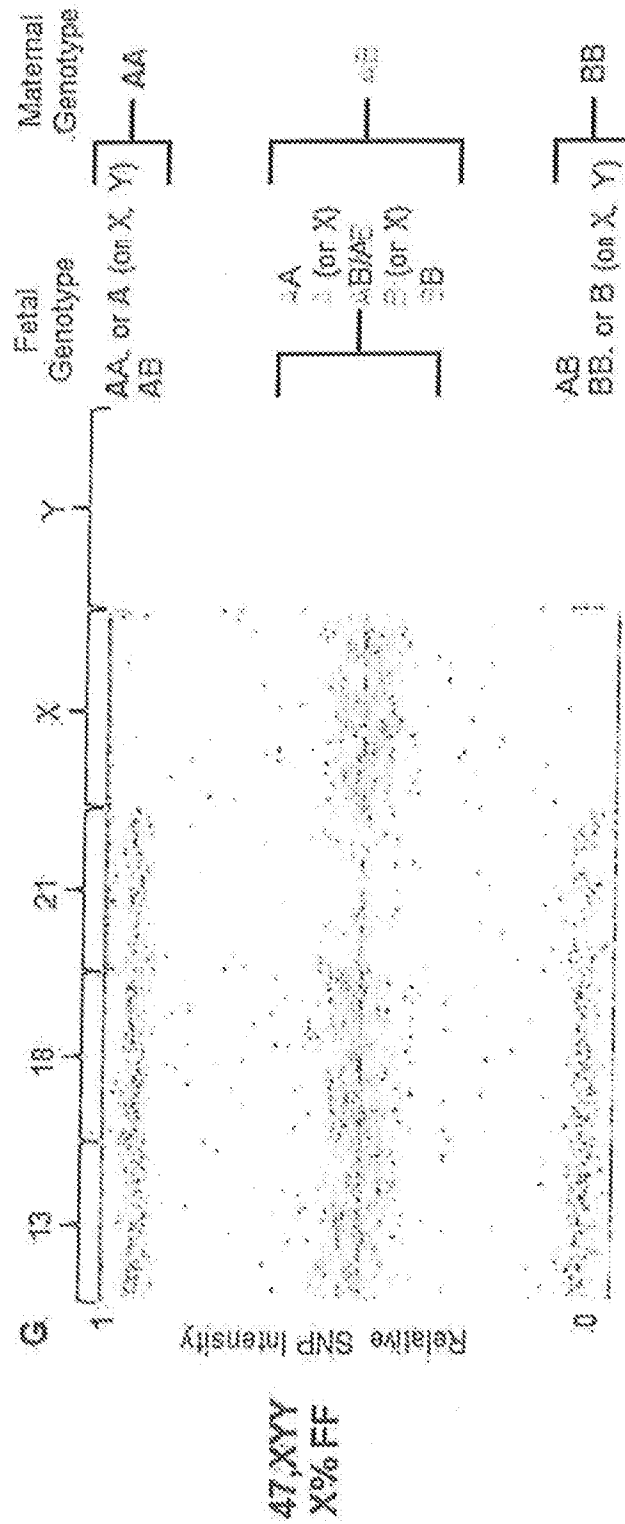


FIG. 31G

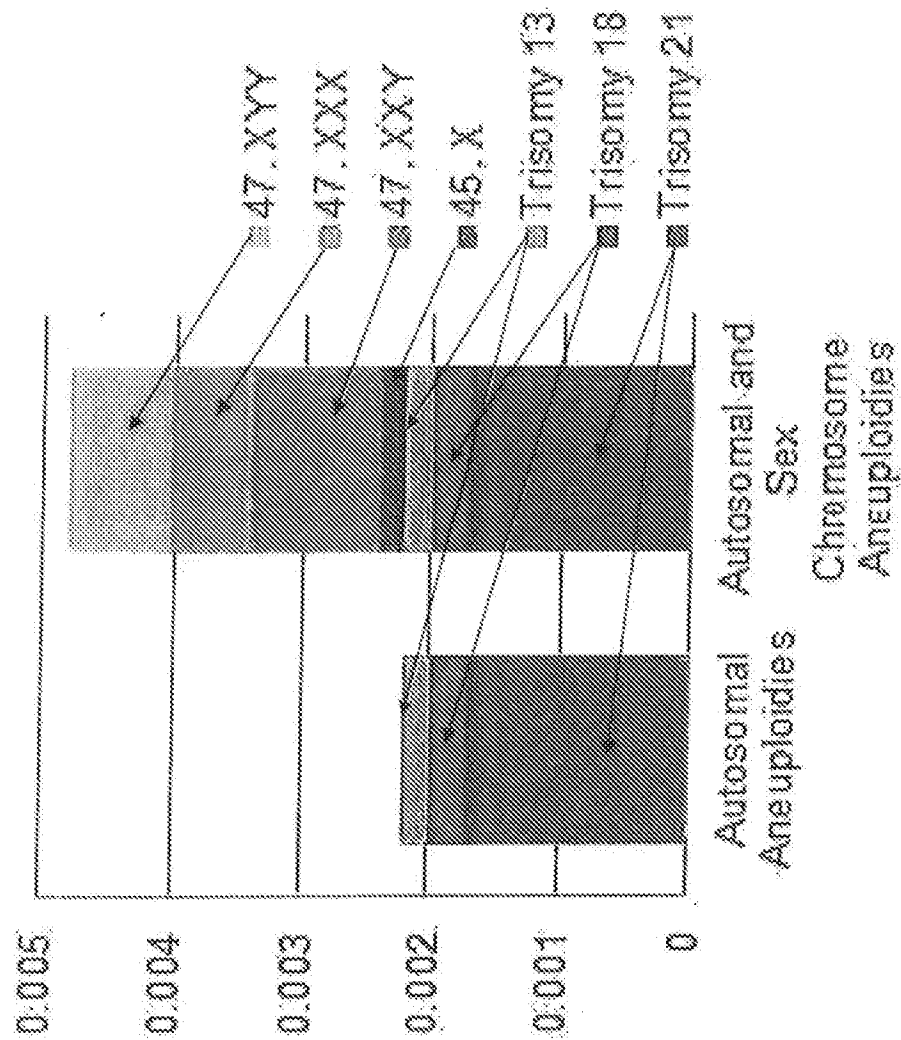


FIG. 32

FIG. 33A

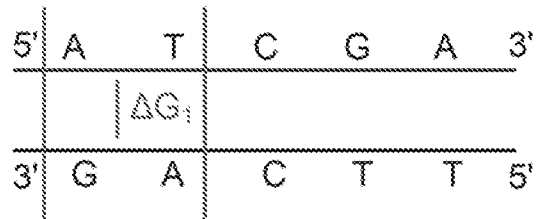
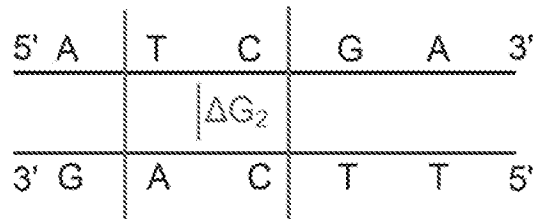


FIG. 33B



◆

◆

◆

$$\Delta G = \Delta G_1 + \Delta G_2 + \dots + \Delta G_x$$

FIG. 33C

5' A T C G A 3'

3' G A C T T 5'

SHIFT
ALIGNMENT
BY 1



FIG. 33D

5' A T C G A 3'

3' G A C T T 5'

FIG. 33E



FIG. 33F

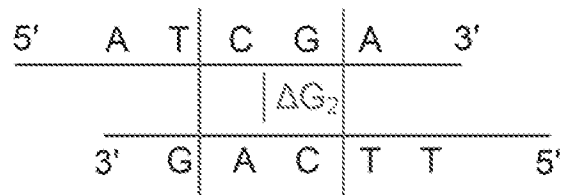


FIG. 34

Description	MAPPED_READS	TOTAL_READS	%Mapped Reads
gMother	3,633,900	3,647,446	100%
gMother	3,567,685	3,578,549	100%
gMother	3,733,706	3,747,692	100%
gChild	3,725,554	3,737,742	100%
gChild	3,816,989	3,829,450	100%
gChild	3,524,956	3,536,569	100%
1 mother	3,467,493	7,564,869	46%
1 mother	3,868,210	7,107,028	54%
1 mother	4,918,140	7,613,240	65%
1 child	361,390	6,507,434	6%
1 child	1,885,864	6,496,348	29%
1 child	2,789,647	6,259,288	45%

FIG. 35

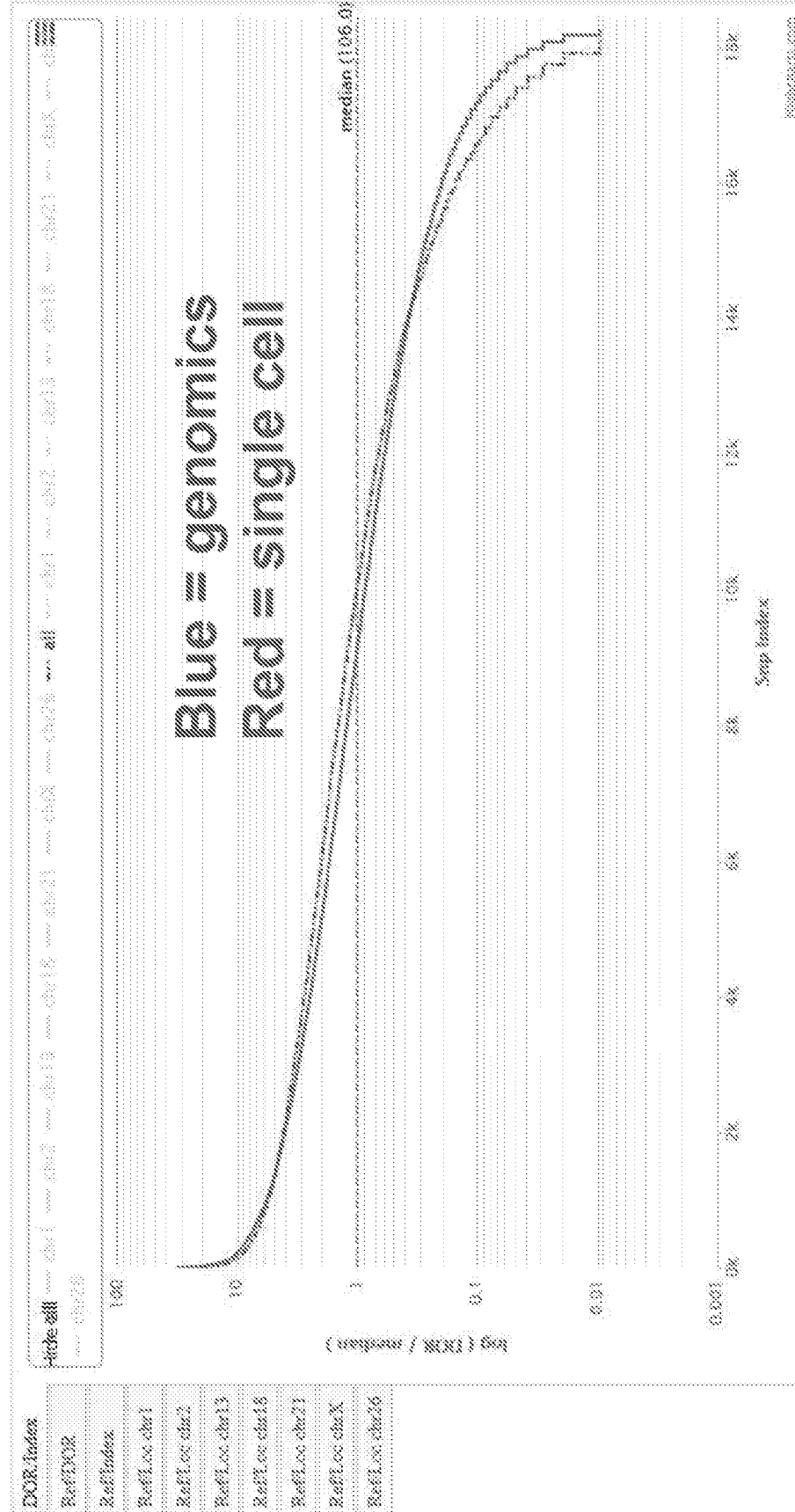


FIG. 36

Description	MAPPED_READS	TOTAL_READS	%Mapped Reads
case 1 blastocoeol	68,123	8,470,872	1%
case 1 cell 1	4,340,613	8,248,598	53%
case 1 cell 2	5,480,580	8,230,870	67%
case 1 cell 3	4,664,577	7,846,040	59%
case 2 blastocoeol	45,794	6,302,957	1%
case 2 cell 1	6,988,854	8,622,495	81%
case 2 cell 2	7,083,600	8,843,495	80%
case 2 cell 3	5,811,364	8,256,310	70%

FIG. 37

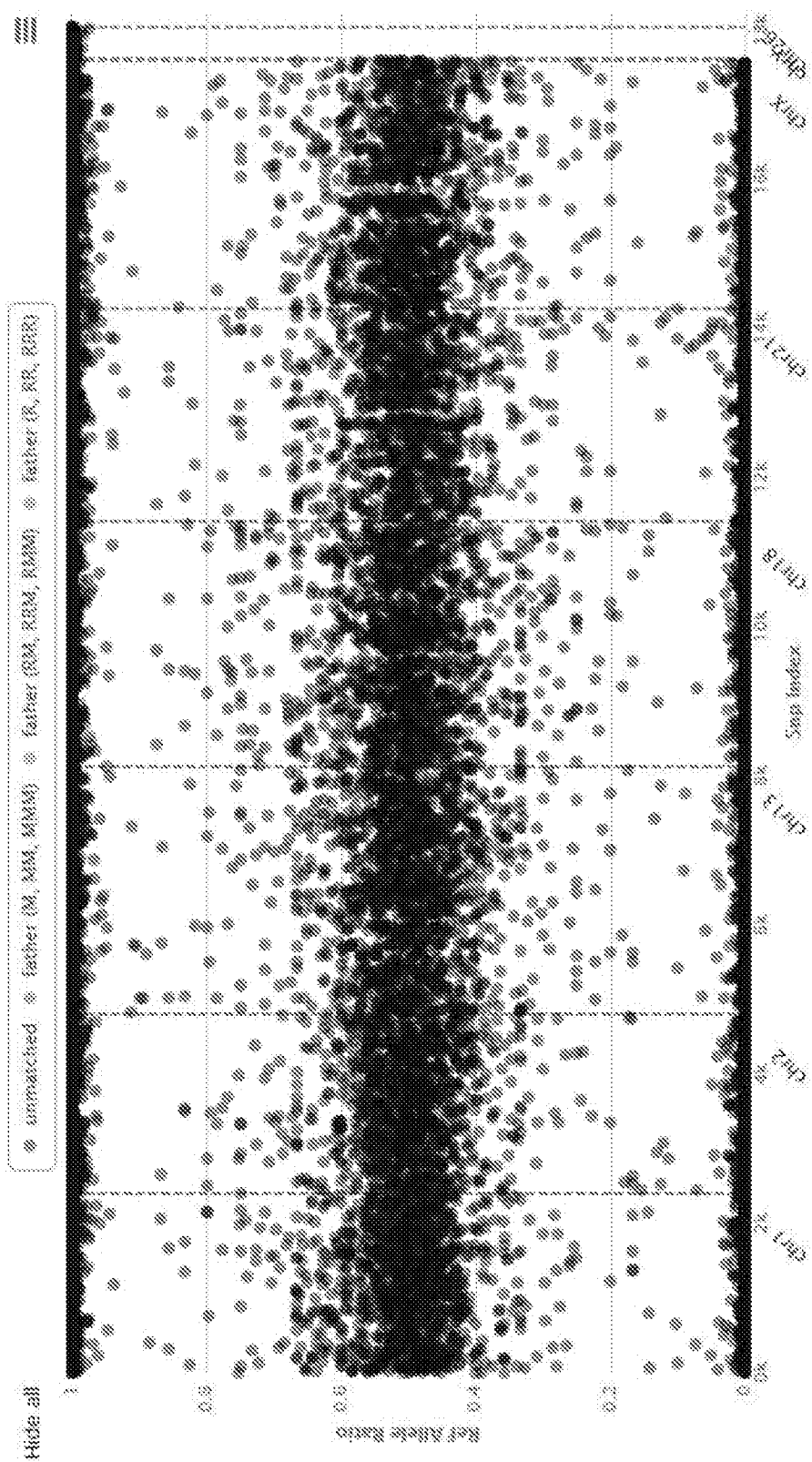


FIG. 38

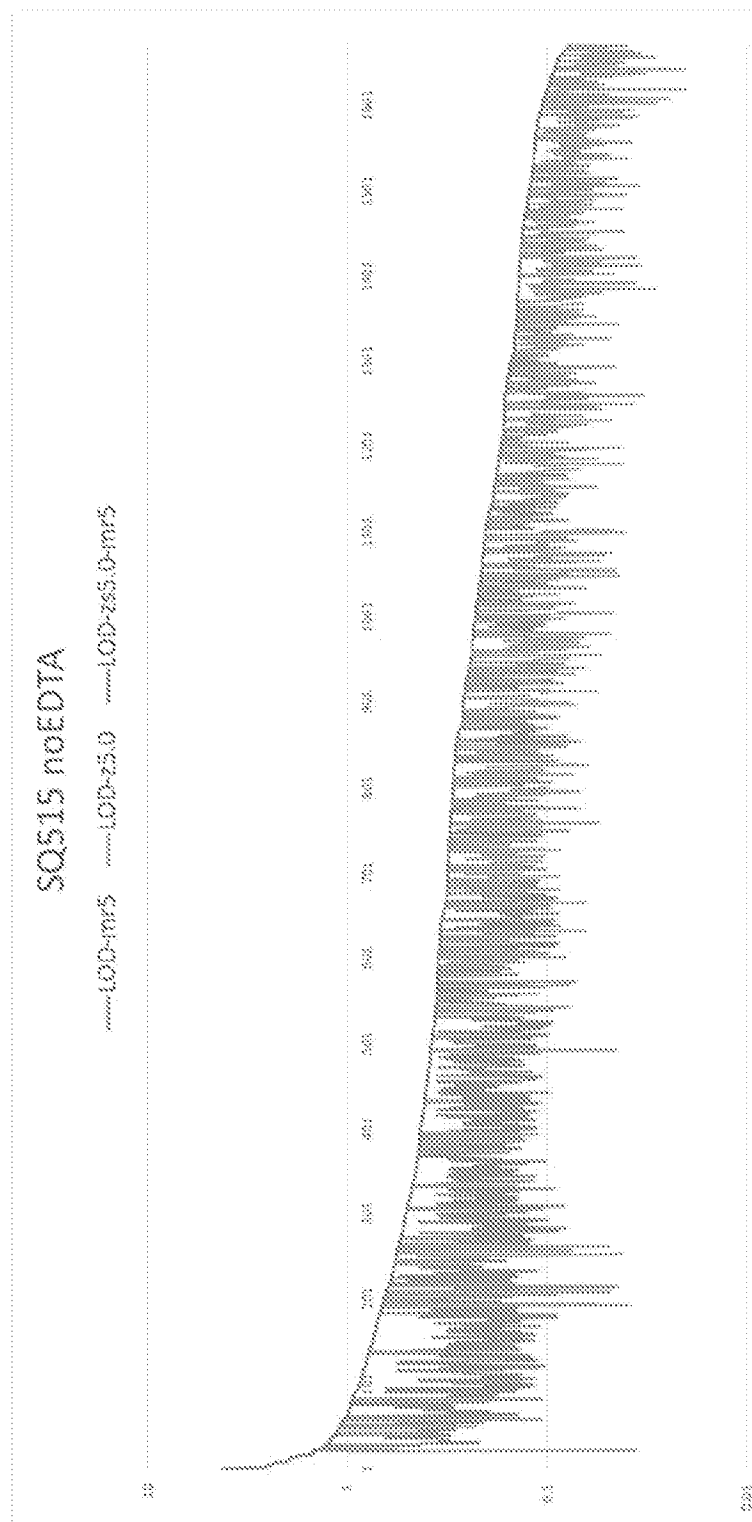
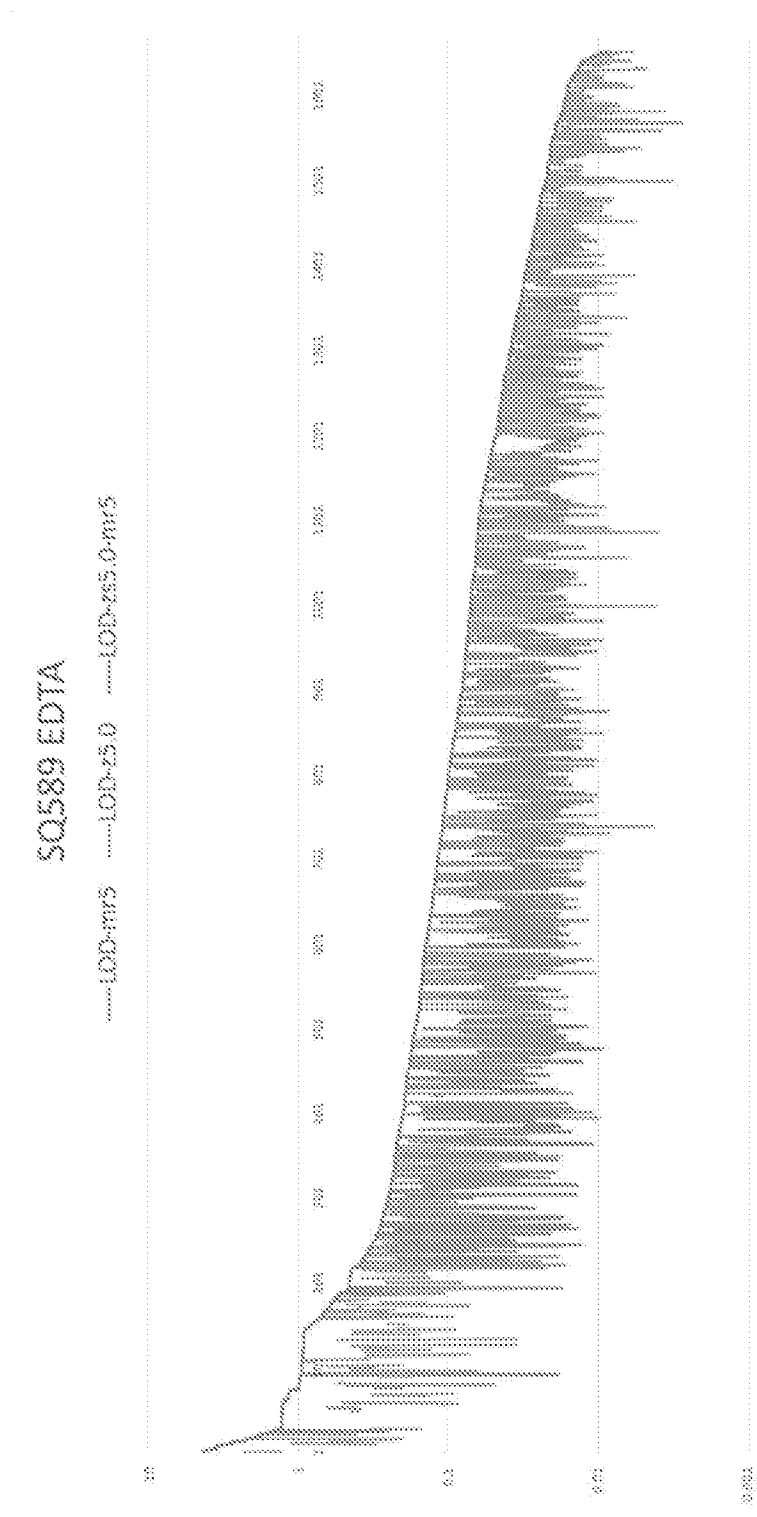
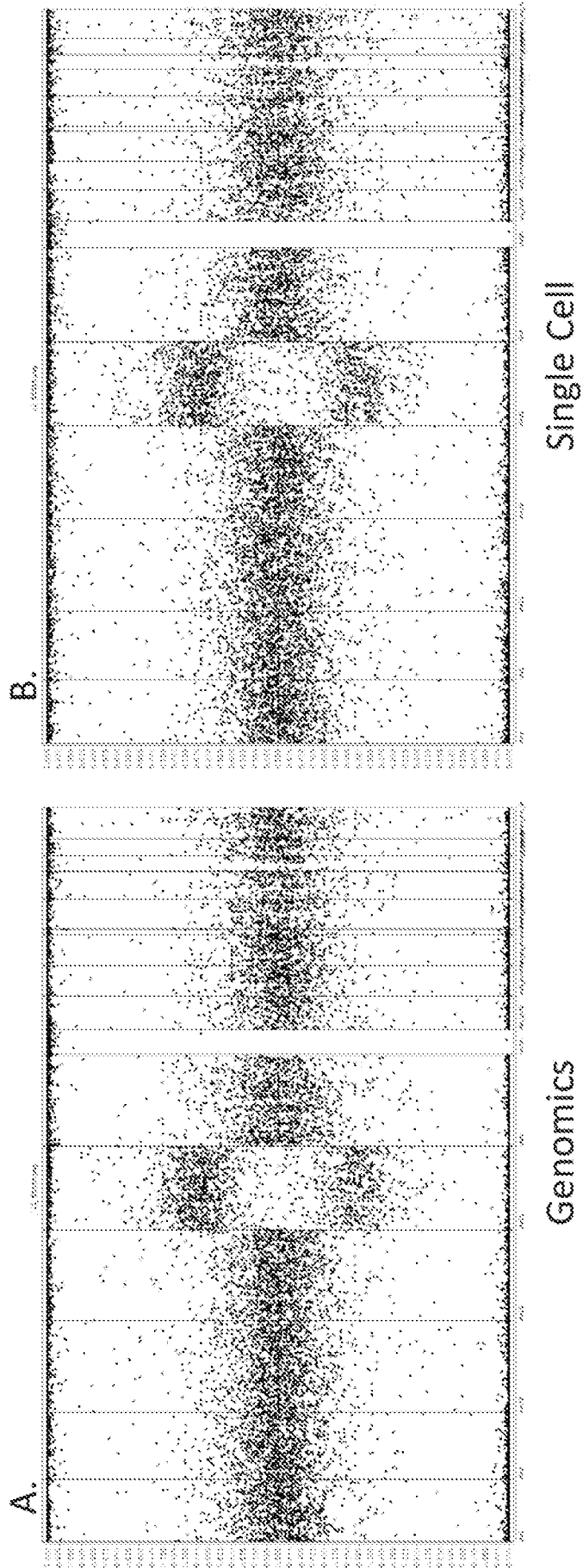


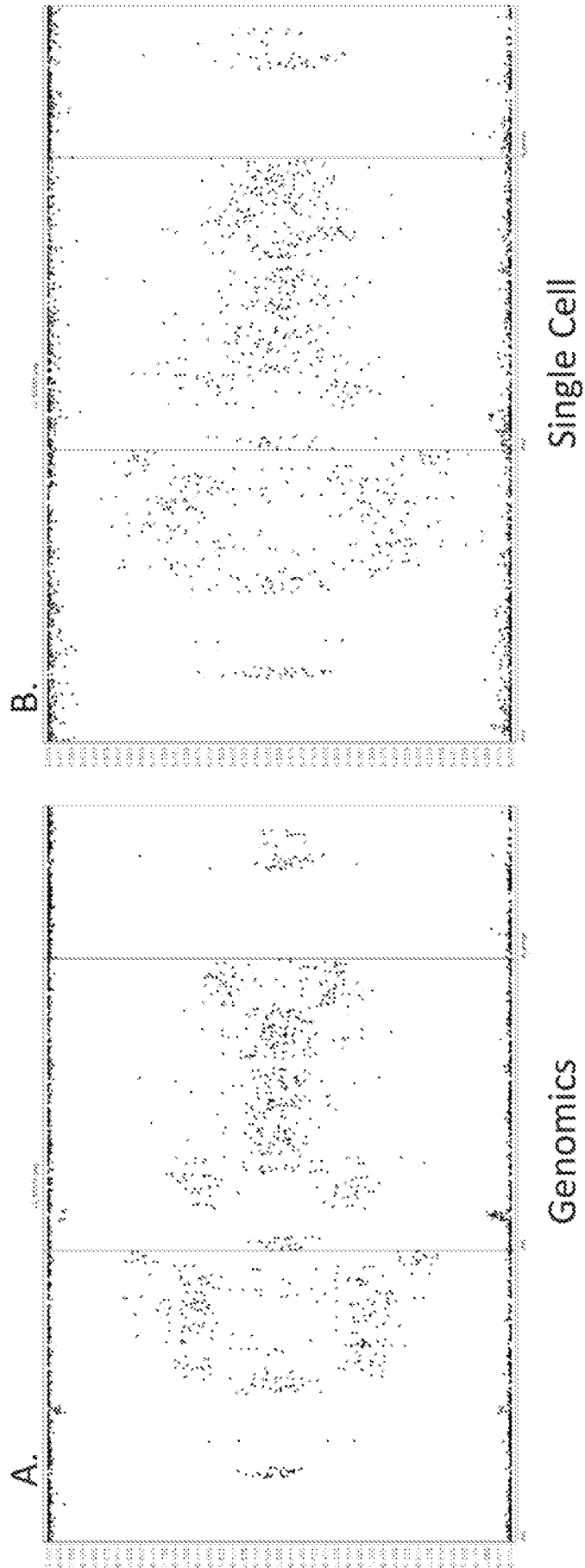
FIG. 39



FIGs. 40A and 40B



FIGs. 41A and 41B



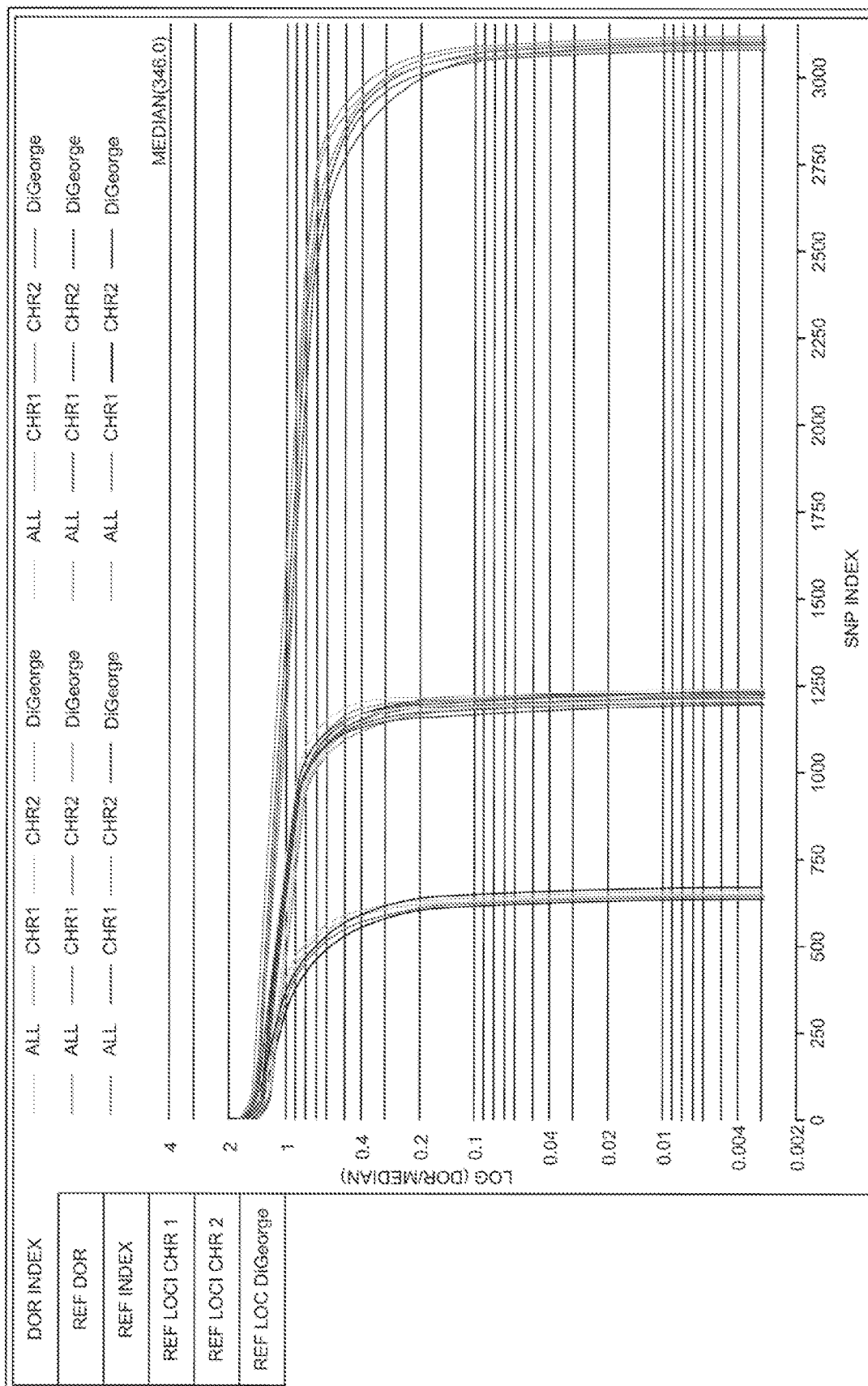


FIG. 42

FIG. 43

	gDNA	Single Cell
Count	75	510
Mean	0.15 %	0.51 %
Median	0.09 %	0.33 %
Max	1.03 %	10 %
Standard Deviation	0.16%	0.79 %
95 th percentile	0.43 %	1.22 %
90 th percentile	0.37 %	0.92 %

FIG. 44

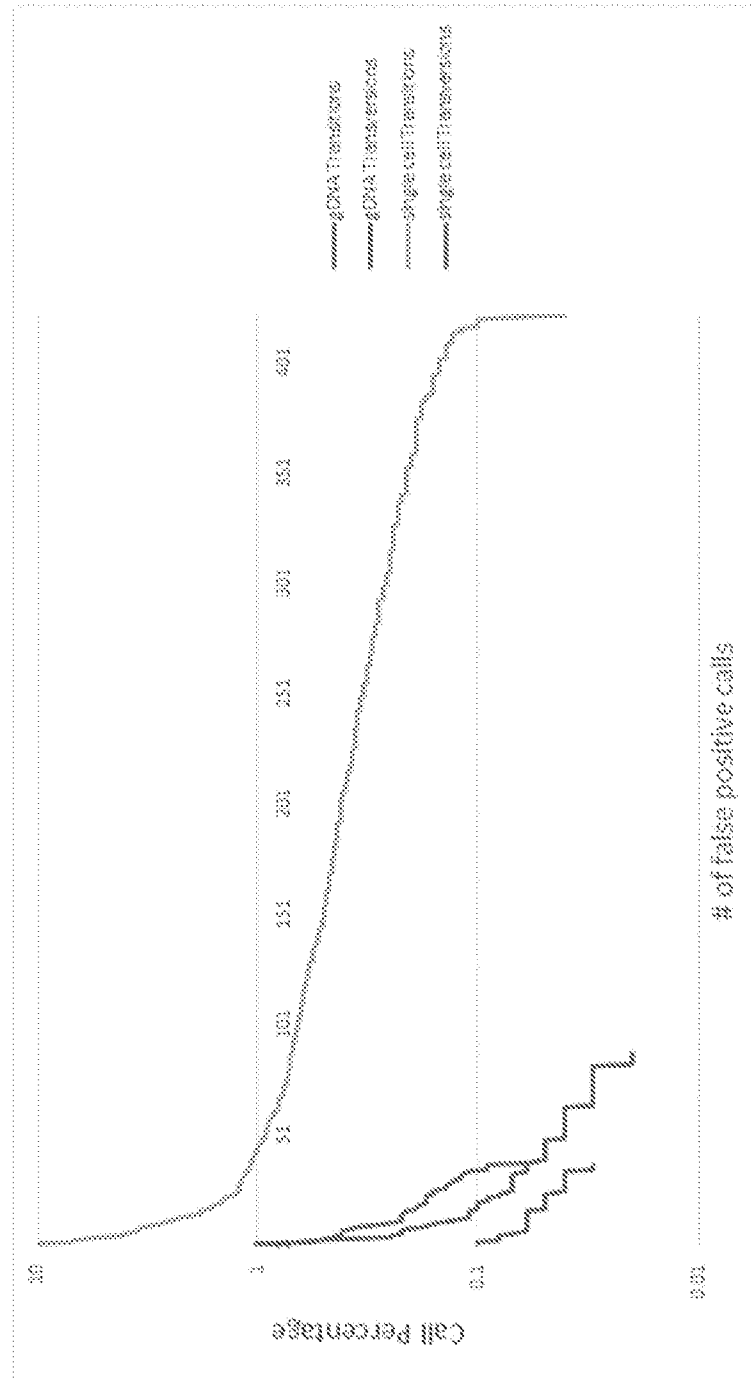


FIG. 45

DOE1	Mapped READS	TOTAL READS	% Mapped	Dropout Count	Median DOR All	NOR	Error Rate (%)
1xMM	1,112,007	1,262,558	88.0%	436	26.25	1,043,539	0.107%
2xMM	3,412,593	3,615,347	94.4%	105	147.75	3,206,478	0.141%
F-A	449,074	635,571	70.6%	1379	9.625	421,228	0.077%
F-B	3,293,944	3,519,378	93.7%	105	142.5	3,106,520	0.072%
F-D	4,028,128	4,109,448	98.0%	122	142.5	3,796,213	0.064%
F-J	3,676,617	3,922,330	93.7%	102	159.25	3,461,304	0.079%
F-A is : 25 mM Tris pH 7.8, 3 mM MgCl ₂ , 0 mM KCl, 10 mM (NH ₄) ₂ SO ₄ , 50 U/mL Taq Gold aka F-A Gold							
F-B is : 75 mM Tris pH 7.8, 6 mM MgCl ₂ , 0 mM KCl, 40 mM (NH ₄) ₂ SO ₄ , 150 U/mL Taq Gold aka F-B Gold							
F-D is : 25 mM Tris pH 8.2 , 3 mM MgCl ₂ , 30 mM KCl, 40 mM (NH ₄) ₂ SO ₄ , 150 U/mL Taq Gold							
F-J is : 75 mM Tris pH 7.8, 6 mM MgCl ₂ , 0 mM KCl, 40 mM (NH ₄) ₂ SO ₄ , 150 U/mL Qiagen HS Taq aka F-B Qiagen							

FIG. 46

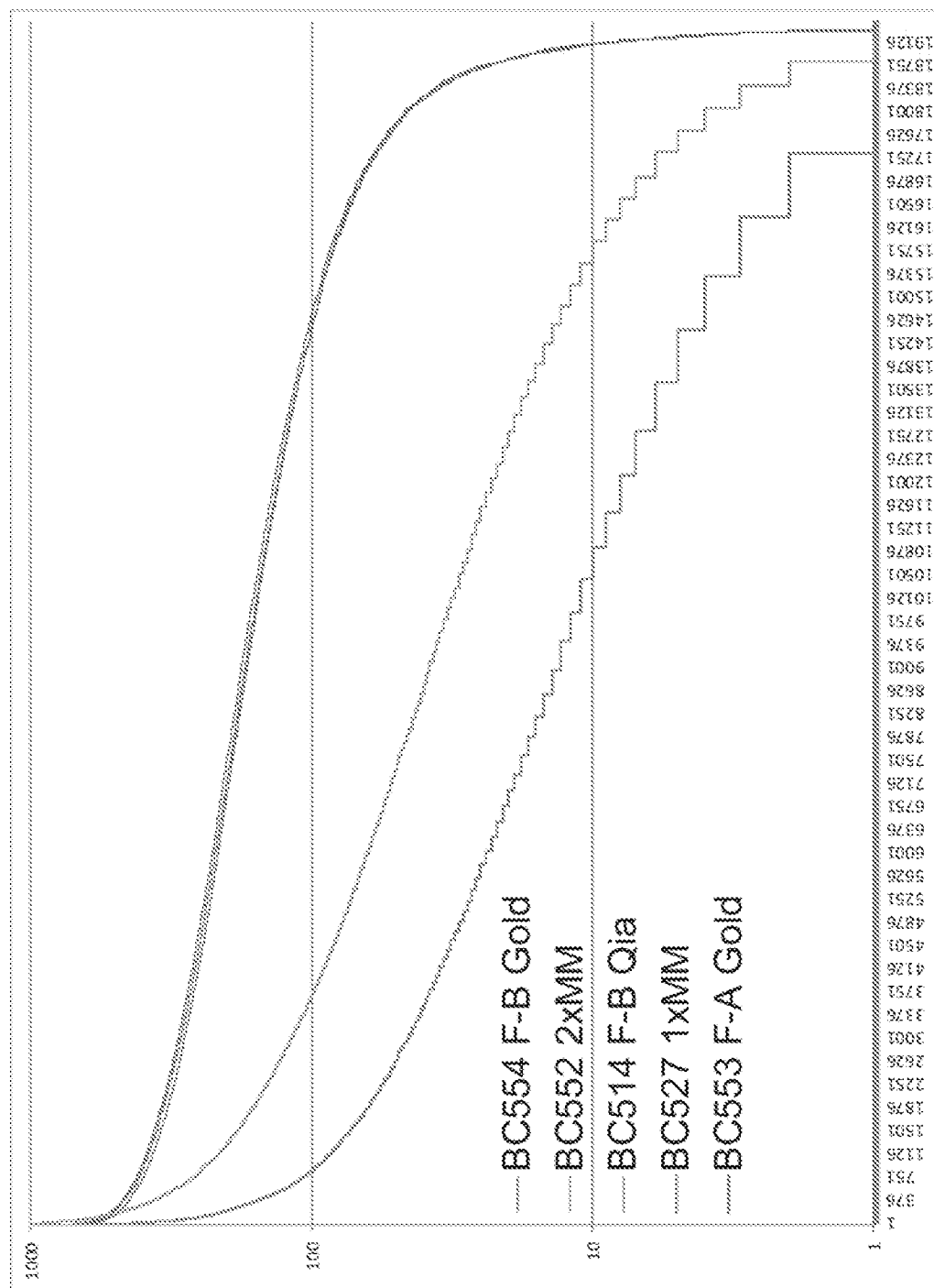
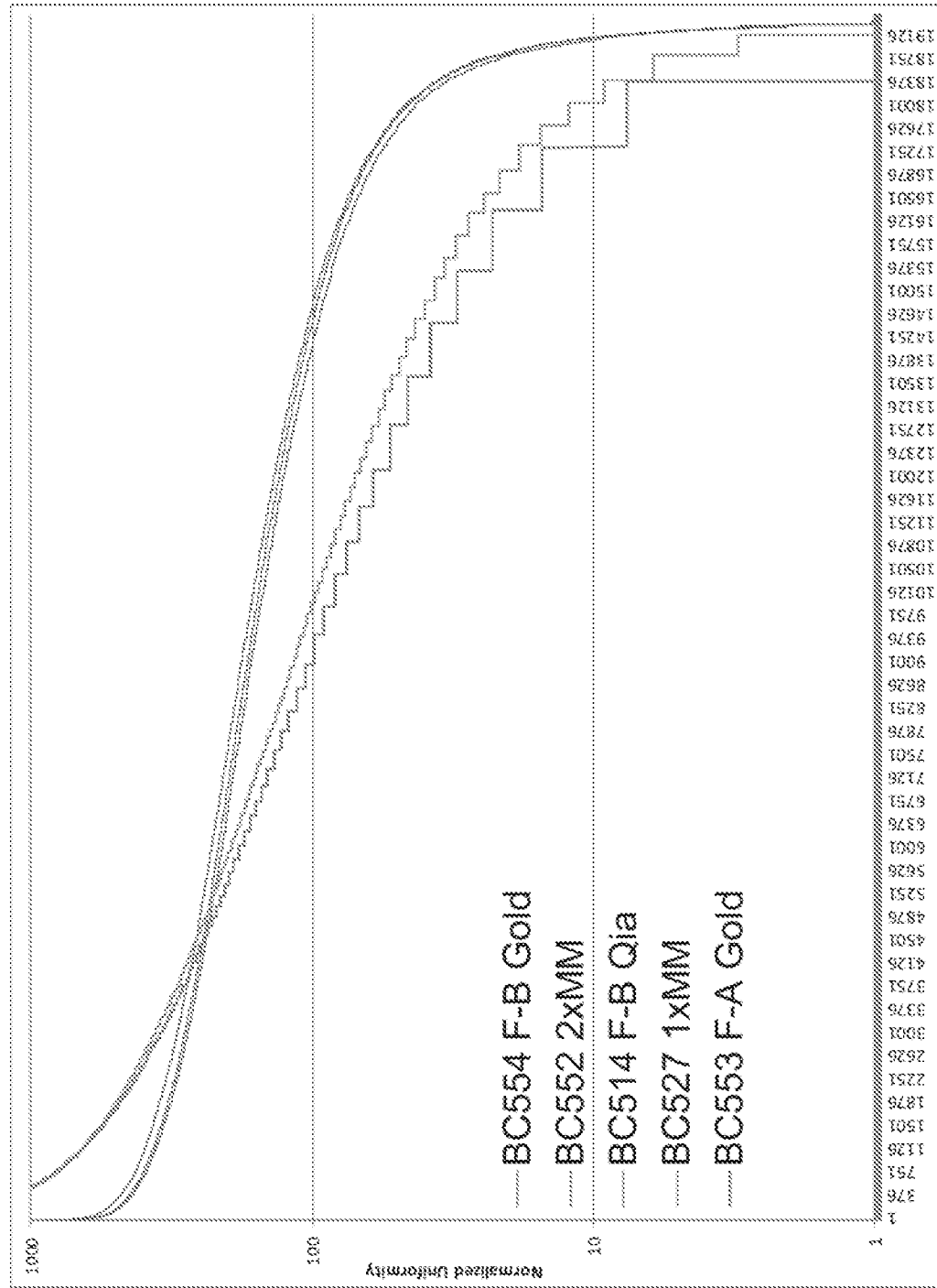


FIG. 47

58/58



Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	N.012.US.05
		Application Number	
Title of Invention	METHODS FOR SIMULTANEOUS AMPLIFICATION OF TARGET LOCI		
The application data sheet is part of the provisional or nonprovisional application for which it is being submitted. The following form contains the bibliographic data arranged in a format specified by the United States Patent and Trademark Office as outlined in 37 CFR 1.76. This document may be completed electronically and submitted to the Office in electronic format using the Electronic Filing System (EFS) or the document may be printed and included in a paper filed application.			

Secrecy Order 37 CFR 5.2

☐	Portions or all of the application associated with this Application Data Sheet may fall under a Secrecy Order pursuant to 37 CFR 5.2 (Paper filers only. Applications that fall under Secrecy Order may not be filed electronically.)
--------------------------	---

Inventor Information:

Inventor 1					Remove
Legal Name					
Prefix	Given Name	Middle Name	Family Name	Suffix	
	Bernhard		Zimmermann		
Residence Information (Select One) ☒ US Residency ☐ Non US Residency ☐ Active US Military Service					
City	San Mateo	State/Province	CA	Country of Residence i	US

Mailing Address of Inventor:

Address 1	1662 Lodi Avenue				
Address 2					
City	San Mateo	State/Province	CA		
Postal Code	94401	Country i	US		

Inventor 2					Remove
Legal Name					
Prefix	Given Name	Middle Name	Family Name	Suffix	
	Matthew	M.	Hill		
Residence Information (Select One) ☒ US Residency ☐ Non US Residency ☐ Active US Military Service					
City	Belmont	State/Province	CA	Country of Residence i	US

Mailing Address of Inventor:

Address 1	2406 Palmer Ave				
Address 2					
City	Belmont	State/Province	CA		
Postal Code	94002	Country i	US		

Inventor 3					Remove
Legal Name					
Prefix	Given Name	Middle Name	Family Name	Suffix	
	Philippe	Gilbert	Lacroute		
Residence Information (Select One) ☒ US Residency ☐ Non US Residency ☐ Active US Military Service					

NeoGenomics

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	N.012.US.05	
		Application Number		
Title of Invention	METHODS FOR SIMULTANEOUS AMPLIFICATION OF TARGET LOCI			

City	Sunnyvale	State/Province	CA	Country of Residence i	US
------	-----------	----------------	----	------------------------	----

Mailing Address of Inventor:

Address 1	1556 Magpie Lane				
Address 2					
City	Sunnyvale	State/Province	CA		
Postal Code	94087	Country i	US		

Inventor 4	Remove
Legal Name	

Prefix	Given Name	Middle Name	Family Name	Suffix
	Michael		Dodd	
Residence Information (Select One) ☒ US Residency ☐ Non US Residency ☐ Active US Military Service				

City	San Francisco	State/Province	CA	Country of Residence i	US
------	---------------	----------------	----	------------------------	----

Mailing Address of Inventor:

Address 1	300 Monterey Blvd.				
Address 2	#203				
City	San Francisco	State/Province	CA		
Postal Code	94131	Country i	US		

All Inventors Must Be Listed - Additional Inventor Information blocks may be generated within this form by selecting the **Add** button.

Add

Correspondence Information:

Enter either Customer Number or complete the Correspondence Information section below.
For further information see 37 CFR 1.33(a).

☐ An Address is being provided for the correspondence Information of this application.

Customer Number	110711		
Email Address		Add Email	Remove Email

Application Information:

Title of the Invention	METHODS FOR SIMULTANEOUS AMPLIFICATION OF TARGET LOCI		
Attorney Docket Number	N.012.US.05	Small Entity Status Claimed	☐
Application Type	Nonprovisional		
Subject Matter	Utility		
Total Number of Drawing Sheets (if any)	55	Suggested Figure for Publication (if any)	

Filing By Reference :

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	N.012.US.05
		Application Number	
Title of Invention	METHODS FOR SIMULTANEOUS AMPLIFICATION OF TARGET LOCI		

Only complete this section when filing an application by reference under 35 U.S.C. 111(c) and 37 CFR 1.57(a). Do not complete this section if application papers including a specification and any drawings are being filed. Any domestic benefit or foreign priority information must be provided in the appropriate section(s) below (i.e., "Domestic Benefit/National Stage Information" and "Foreign Priority Information").

For the purposes of a filing date under 37 CFR 1.53(b), the description and any drawings of the present application are replaced by this reference to the previously filed application, subject to conditions and requirements of 37 CFR 1.57(a).

Application number of the previously filed application	Filing date (YYYY-MM-DD)	Intellectual Property Authority or Country

Publication Information:

☐ Request Early Publication (Fee required at time of Request 37 CFR 1.219)

☐ **Request Not to Publish.** I hereby request that the attached application not be published under 35 U.S.C. 122(b) and certify that the invention disclosed in the attached application **has not and will not** be the subject of an application filed in another country, or under a multilateral international agreement, that requires publication at eighteen months after filing.

Representative Information:

Representative information should be provided for all practitioners having a power of attorney in the application. Providing this information in the Application Data Sheet does not constitute a power of attorney in the application (see 37 CFR 1.32). Either enter Customer Number or complete the Representative Name section below. If both sections are completed the customer Number will be used for the Representative Information during processing.

Please Select One:	☒ Customer Number	☐ US Patent Practitioner	☐ Limited Recognition (37 CFR 11.9)
Customer Number	110711		

Domestic Benefit/National Stage Information:

This section allows for the applicant to either claim benefit under 35 U.S.C. 119(e), 120, 121, or 365(c) or indicate National Stage entry from a PCT application. Providing this information in the application data sheet constitutes the specific reference required by 35 U.S.C. 119(e) or 120, and 37 CFR 1.78.

When referring to the current application, please leave the application number blank.

Prior Application Status	Pending	Remove	
Application Number	Continuity Type	Prior Application Number	Filing Date (YYYY-MM-DD)
	Claims benefit of provisional	62066514	2014-10-21
Prior Application Status	Pending	Remove	
Application Number	Continuity Type	Prior Application Number	Filing Date (YYYY-MM-DD)
	Claims benefit of provisional	62146188	2015-04-10
Prior Application Status	Pending	Remove	

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	N.012.US.05
		Application Number	
Title of Invention	METHODS FOR SIMULTANEOUS AMPLIFICATION OF TARGET LOCI		

Application Number	Continuity Type	Prior Application Number	Filing Date (YYYY-MM-DD)
	Claims benefit of provisional	62147377	2015-04-14
Prior Application Status	Pending	Remove	
Application Number	Continuity Type	Prior Application Number	Filing Date (YYYY-MM-DD)
	Claims benefit of provisional	62148173	2015-04-15
Prior Application Status	Pending	Remove	
Application Number	Continuity Type	Prior Application Number	Filing Date (YYYY-MM-DD)
	Continuation in part of	14538982	2014-11-24
Prior Application Status	Pending	Remove	
Application Number	Continuity Type	Prior Application Number	Filing Date (YYYY-MM-DD)
	Continuation in part of	14692703	2015-04-21
Prior Application Status	Pending	Remove	
Application Number	Continuity Type	Prior Application Number	Filing Date (YYYY-MM-DD)
	Continuation in part of	14877925	2015-10-07
Additional Domestic Benefit/National Stage Data may be generated within this form by selecting the Add button. Add			

Foreign Priority Information:

This section allows for the applicant to claim priority to a foreign application. Providing this information in the application data sheet constitutes the claim for priority as required by 35 U.S.C. 119(b) and 37 CFR 1.55(d). When priority is claimed to a foreign application that is eligible for retrieval under the priority document exchange program (PDX) the information will be used by the Office to automatically attempt retrieval pursuant to 37 CFR 1.55(h)(1) and (2). Under the PDX program, applicant bears the ultimate responsibility for ensuring that a copy of the foreign application is received by the Office from the participating foreign intellectual property office, or a certified copy of the foreign priority application is filed, within the time period specified in 37 CFR 1.55(g)(1).

Remove			
Application Number	Country ⁱ	Filing Date (YYYY-MM-DD)	Access Code ^j (if applicable)
Additional Foreign Priority Data may be generated within this form by selecting the Add button. Add			

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	N.012.US.05
		Application Number	
Title of Invention	METHODS FOR SIMULTANEOUS AMPLIFICATION OF TARGET LOCI		

Statement under 37 CFR 1.55 or 1.78 for AIA (First Inventor to File) Transition Applications

This application (1) claims priority to or the benefit of an application filed before March 16, 2013 and (2) also contains, or contained at any time, a claim to a claimed invention that has an effective filing date on or after March 16, 2013. ☐ NOTE: By providing this statement under 37 CFR 1.55 or 1.78, this application, with a filing date on or after March 16, 2013, will be examined under the first inventor to file provisions of the AIA.
--

Authorization to Permit Access:

☒ Authorization to Permit Access to the Instant Application by the Participating Offices
If checked, the undersigned hereby grants the USPTO authority to provide the European Patent Office (EPO), the Japan Patent Office (JPO), the Korean Intellectual Property Office (KIPO), the World Intellectual Property Office (WIPO), and any other intellectual property offices in which a foreign application claiming priority to the instant patent application is filed access to the instant patent application. See 37 CFR 1.14(c) and (h). This box should not be checked if the applicant does not wish the EPO, JPO, KIPO, WIPO, or other intellectual property office in which a foreign application claiming priority to the instant patent application is filed to have access to the instant patent application. In accordance with 37 CFR 1.14(h)(3), access will be provided to a copy of the instant patent application with respect to: 1) the instant patent application-as-filed; 2) any foreign application to which the instant patent application claims priority under 35 U.S.C. 119(a)-(d) if a copy of the foreign application that satisfies the certified copy requirement of 37 CFR 1.55 has been filed in the instant patent application; and 3) any U.S. application-as-filed from which benefit is sought in the instant patent application. In accordance with 37 CFR 1.14(c), access may be provided to information concerning the date of filing this Authorization.

Applicant Information:

Providing assignment information in this section does not substitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.
--

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	N.012.US.05
		Application Number	
Title of Invention	METHODS FOR SIMULTANEOUS AMPLIFICATION OF TARGET LOCI		

Applicant 1				Remove	
If the applicant is the inventor (or the remaining joint inventor or inventors under 37 CFR 1.45), this section should not be completed. The information to be provided in this section is the name and address of the legal representative who is the applicant under 37 CFR 1.43; or the name and address of the assignee, person to whom the inventor is under an obligation to assign the invention, or person who otherwise shows sufficient proprietary interest in the matter who is the applicant under 37 CFR 1.46. If the applicant is an applicant under 37 CFR 1.46 (assignee, person to whom the inventor is obligated to assign, or person who otherwise shows sufficient proprietary interest) together with one or more joint inventors, then the joint inventor or inventors who are also the applicant should be identified in this section.					
Clear					
☒ Assignee		☐ Legal Representative under 35 U.S.C. 117		☐ Joint Inventor	
☐ Person to whom the inventor is obligated to assign.			☐ Person who shows sufficient proprietary interest		
If applicant is the legal representative, indicate the authority to file the patent application, the inventor is:					
Name of the Deceased or Legally Incapacitated Inventor :					
If the Applicant is an Organization check here. ☒					
Organization Name		Natera, Inc.			
Mailing Address Information:					
Address 1		201 Industrial Road			
Address 2		Suite 410			
City		San Carlos		State/Province	CA
Country ⁱ	US	Postal Code		94070	
Phone Number				Fax Number	
Email Address		IP@natera.com			
Additional Applicant Data may be generated within this form by selecting the Add button. Add					

Assignee Information including Non-Applicant Assignee Information:

Providing assignment information in this section does not substitute for compliance with any requirement of part 3 of Title 37 of CFR to have an assignment recorded by the Office.	
Assignee 1	
Complete this section if assignee information, including non-applicant assignee information, is desired to be included on the patent application publication. An assignee-applicant identified in the "Applicant Information" section will appear on the patent application publication as an applicant. For an assignee-applicant, complete this section only if identification as an assignee is also desired on the patent application publication.	
Remove	
If the Assignee or Non-Applicant Assignee is an Organization check here. ☒	

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Application Data Sheet 37 CFR 1.76		Attorney Docket Number	N.012.US.05
		Application Number	
Title of Invention	METHODS FOR SIMULTANEOUS AMPLIFICATION OF TARGET LOCI		

Organization Name	Natera, Inc.
-------------------	--------------

Mailing Address Information For Assignee including Non-Applicant Assignee:

Address 1	201 Industrial Road		
Address 2	Suite 410		
City	San Carlos	State/Province	CA
Country i	US	Postal Code	94070
Phone Number		Fax Number	
Email Address	IP@natera.com		

Additional Assignee or Non-Applicant Assignee Data may be generated within this form by selecting the Add button.

Signature:

NOTE: This form must be signed in accordance with 37 CFR 1.33. See 37 CFR 1.4 for signature requirements and certifications

Signature	/Anton Bokal/		Date (YYYY-MM-DD)	2015-10-20	
First Name	Anton	Last Name	Bokal	Registration Number	51243

Additional Signature may be generated within this form by selecting the Add button.

This collection of information is required by 37 CFR 1.76. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 23 minutes to complete, including gathering, preparing, and submitting the completed application data sheet form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C. 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether the Freedom of Information Act requires disclosure of these records.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspections or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Appl. No.:	Not Yet Assigned	Confirmation No.: Not Yet Assigned
First Named Inventor:	Bernhard Zimmermann	Art Unit: 1637
Filed:	Herewith	Examiner: Mummert, Stephanie Kane
For:	METHODS FOR SIMULTANEOUS AMPLIFICATION OF TARGET LOCI	Docket No.: N.012.US.05
Customer No.:	110711	

PRELIMINARY AMENDMENT

COMMISSIONER FOR PATENTS
U.S. Patent and Trademark Office
Via EFS-Web

Dear Commissioner:

Applicant kindly requests amendment of the claims as provided herein prior to examination on the merits.

- ☒ **Amendments to the Claims** begin on page 2 of this paper.
- ☒ **Remarks** begin on page 5 of this paper.

Amendments to the Claims:

The following listing of claims will replace all prior versions and listings of claims in the application.

Listing of Claims:

1. (original) A method of sequencing target loci in a nucleic acid sample, the method comprising:

(a) contacting in one reaction volume, a nucleic acid sample comprising DNA molecules comprising target human loci and a library of at least 1,000 non-immobilized, non-identical primers that hybridize to at least 1,000 non-identical target human loci to produce a reaction mixture in solution in which the concentration of each primer is less than 20 nM; wherein the primers do not include molecular inversion probes (MIPs); and wherein the average length of the DNA molecules is less than 200 base pairs;

(b) subjecting the reaction mixture to primer extension reaction conditions that include an annealing step to produce amplified products comprising target amplicons; wherein the length of the annealing step for the reaction conditions is greater than 10 minutes; wherein the length of the target amplicons is less than 100 nucleotides; wherein at least 1,000 different target human loci are amplified, and wherein at least 50% of the amplified products comprise target loci; and

(c) sequencing the amplified human loci, wherein the method does not comprise using a microarray, thereby sequencing target loci in a nucleic acid sample.

2. (New) The method of claim 1, wherein the range of melting temperatures of the at least 1000 non-identical primers is less than 10 °C.

3. (New) The method of claim 1, wherein the range of melting temperatures of the at least 1000 non-identical primers is less than 5 °C.

4. (New) The method of claim 1, further comprising performing universal amplification on DNA molecules in the sample prior to step (a).

5. (New) The method of claim 1, wherein at least 90% of the amplified products are target amplicons.
6. (New) The method of claim 1, wherein at least 90% of the target loci are amplified.
7. (New) The method of claim 1, wherein less than 20% of the amplified products are primer dimers.
8. (New) The method of claim 1, wherein the at least 1,000 non-identical primers comprise a 5' region that is not specific for a target human locus followed by a region that is specific for a target human locus, an internal region that is not specific for the target human locus and forms a loop structure, and a 3' region that is specific for the target human locus.
9. (New) The method of claim 1, wherein the sample comprises maternal DNA from a pregnant mother of a fetus and fetal DNA, and wherein the method further comprises determining the presence or absence of a fetal chromosome abnormality from the sequencing data.
10. (New) The method of claim 1, wherein the sample comprises DNA from a single cell.
11. (New) The method of claim 1, wherein the sample comprises blood.
12. (New) The method of claim 11, wherein the method further comprises, making a genetic diagnosis, wherein the genetic diagnosis comprises detecting a gene mutation or detecting a ploidy state of the target human loci.
13. (New) The method of claim 1, wherein the length of the annealing step of the reaction conditions is between 10 and 60 minutes.
14. (New) The method of claim 1, wherein the concentration of primers in the reaction volume are less than 10 nM.

15. (New) The method of claim 1, wherein the concentration of primers in the reaction volume are less than 5 nM.

16. (New) The method of claim 1, wherein the library comprises at least 7,5000 non-immobilized, non-identical primers that hybridize to at least 7,500 non-identical target human loci, and wherein the 7,500 non-identical target human loci are amplified in step c.

17. (New) The method of claim 1, The library of claim 1, wherein a range of guanine-cytosine (GC) content among the different test primers in the library is less than 30%.

18. (New) The method of claim 15, wherein a range of melting temperatures among the different primers in the library is less than 20 °C.

19. (New) The method of claim 16, wherein a range of the length among the different target amplicons is less than 50 nucleotides.

20. (New) The method of claim 1, wherein the target human loci comprise single nucleotide polymorphisms.

Remarks/Arguments

I. Status of the Claims:

Upon entry of this amendment, claims 1-20 are pending. Applicant respectfully reserves the right to present other claims in one or more continuation applications.

Conclusion

Applicant does not believe that any additional fees are due in connection with this amendment. However, in the unlikely event that any such fees are due, the Commissioner is hereby authorized to charge the same to Deposit Account No. 50-6075 (N.012.US.05).

If the Examiner believes for any reason that personal communication will expedite prosecution of this application, the Examiner is invited to telephone the undersigned at the number provided below.

Respectfully submitted,

Date: October 20, 2015

/Anton Bokal/
Anton Bokal
Registration No. 51243
Customer No. 110711
IP@natera.com
650-720-5573
Attorney for Applicant