# MATERNAL PLASMA TRANSCRIPTOME ANALYSIS BY MASSIVELY PARALLEL RNA SEQUENCING

## CROSS-REFERENCES TO RELATED APPLICATIONS

**[0001]** The present application claims priority to U.S. Provisional Patent Application No. 61/770,985, entitled "MATERNAL PLASMA TRANSCRIPTOME ANALYSIS BY MASSIVELY PARALLEL RNA SEQUENCING" and filed February 28, 2013, which is incorporated herein by reference for all purposes.

## BACKGROUND

**[0002]** Extracellular RNA molecules that are pregnancy-specific in maternal plasma have been reported [1]. They have provided a noninvasive testing tool for fetal assessment and pregnancy monitoring by simply using a peripheral blood specimen from the mother. To date, a number of promising prenatal diagnostic applications have been developed using maternal plasma RNA [2-8]. Researchers have been actively searching for additional RNA markers in order to extend the applications in different contexts of fetal disorders and pregnancy pathologies.

**[0003]** Theoretically, the most straightforward RNA marker identification method is to directly profile extracellular RNA molecules in maternal plasma. This method has not however been easy because conventional high-throughput screening technologies, such as microarray analysis and serial analysis of gene expression (SAGE), have a limited ability to detect the typically low concentrations of partially degraded extracellular RNA in maternal plasma [9]. Instead, most of the reported RNA marker screening strategies operate indirectly by comparing the expression profiles of the placenta and the maternal blood cells [10]. Only the transcripts that are expressed much higher in the placenta than in the maternal blood cells are further studied in maternal plasma samples by high sensitivity but low-throughput technologies, such as real-time reverse transcriptase polymerase chain reaction (RT-PCR). Thus far plasma RNA markers identified by this indirect method are relatively limited. This is possibly because the tissue-based mining strategy has not fully taken into account all of the

biological factors influencing placental RNA levels in maternal plasma. In addition, transcripts that are expressed and released by non-placental tissues in response to pregnancy could not be identified by this method. A sensitive and high-throughput methodology that allows the direct profiling of the maternal plasma transcriptome would therefore be very desirable.

**[0004]** Similarly, direct plasma RNA profiling may be useful in other scenarios where there is a mixture of RNA molecules from two individuals, such as for organ transplantation. The circulation of transplantation recipients contains nucleic acid molecules from both the donor and recipient. A change in the relative profile of RNA molecules contributed by the donor or the recipient may reveal pathologies in the transplanted organ or recipient, such as graft rejection.

## BRIEF SUMMARY

In one aspect, the present invention provides a method of diagnosing a pregnancy-associated disorder using a sample from a female subject pregnant with a fetus, the method comprising:

receiving a plurality of reads, wherein the reads are obtained from an analysis of RNA molecules obtained from the sample, the sample containing a mixture of maternal- and fetal-derived RNA molecules;

identifying, by a computer system, locations of the reads in a reference sequence;

identifying one or more informative loci, each of which is homozygous in a first entity for a corresponding first allele and which is heterozygous in a second entity for the corresponding first allele and a corresponding second allele, wherein the first entity is the pregnant female subject or the fetus, and the second entity is the other one of the pregnant female subject and the fetus;

filtering the one or more informative loci to identify one or more filtered informative loci:

that are located within an expressed region of the reference sequence,

at which at least a first predetermined number of reads in the plurality of reads containing the corresponding first allele are located, and

at which at least a second predetermined number of reads in the plurality of reads containing the corresponding second allele are located,

for each of the filtered informative loci:

determining a first number of reads located at the filtered informative locus and containing the corresponding first allele,

determining a second number of reads located at the filtered informative locus and containing the corresponding second allele,

calculating a first sum of the first numbers and the second sum of the second numbers;

calculating a ratio of the first sum and the second sum, and

comparing the ratio to a cutoff value to determine whether the fetus, mother, or pregnancy has a pregnancy-associated disorder, the cutoff value being determined from one or more samples obtained from pregnant female subjects without the pregnancy-associated disorder.

In another aspect, the present invention provides a method of determining, in a sample from a female subject pregnant with a fetus, a portion of RNA that is of fetal origin, the method comprising:

receiving a plurality of reads, wherein the reads are obtained from an analysis of RNA molecules obtained from the sample, the sample containing a mixture of maternal- and fetal-derived RNA molecules;

identifying, by a computer system, locations of the reads in a reference sequence;

identifying one or more informative maternal loci, each of which is homozygous in the fetus for a corresponding first allele and which is heterozygous in the pregnant female subject for the corresponding first allele and a corresponding second allele;

filtering the one or more informative maternal loci to identify one or more filtered informative maternal loci:

that are located within an expressed region of the reference sequence,

at which at least a first predetermined number of reads in the plurality of reads containing the corresponding first allele are located, and

at which at least a second predetermined number of reads in the plurality of reads containing the corresponding second allele are located,

for each of the filtered informative maternal loci:

determining a first number of reads located at the filtered informative maternal locus and containing the corresponding first allele,

determining a second number of reads located at the filtered informative maternal locus and containing the corresponding second allele,

calculating a sum of the first number and the second number,

calculating a maternal ratio of the second number divided by the sum,

determining a scalar representing a total expression at the filtered informative maternal locus relative to expression of the corresponding second allele in the pregnant female subject,

multiplying the maternal ratio by the scalar to obtain a maternal contribution,

calculating a fetal contribution as one minus the maternal contribution, and determining a portion of RNA in the sample that is of fetal origin, wherein the portion is an average of the fetal contributions for the filtered informative maternal loci.

In yet another aspect, the present invention provides a method of designating a genomic locus as a maternal or fetal marker by analyzing a sample from a female subject pregnant with a fetus, the method comprising:

receiving a plurality of reads, wherein the reads are obtained from an analysis of RNA molecules obtained from the sample, the sample containing a mixture of maternal- and fetal-derived RNA molecules;

identifying, by a computer system, locations of the reads in a reference sequence;

identifying one or more informative loci, each of which is homozygous in a first entity for a corresponding first allele and which is heterozygous in a second entity for the corresponding first allele and a corresponding second allele, wherein the first entity is the pregnant female subject or the fetus, and the second entity is the other one of the pregnant female subject and the fetus;

filtering the one or more informative loci to identify one or more filtered informative loci:

that are located within an expressed region of the reference sequence,

at which at least a first predetermined number of reads in the plurality of reads containing the corresponding first allele are located, and

at which at least a second predetermined number of reads in the plurality of reads containing the corresponding second allele are located,

for each of the filtered informative loci:

determining a first number of reads located at the filtered informative locus and containing the corresponding first allele,

determining a second number of reads located at the filtered informative locus and containing the corresponding second allele,

calculating a ratio of the first number and the second number,

designating the filtered informative locus as a marker for the second entity when the ratio exceeds a cutoff.

In yet another aspect, the present invention provides use of a gene listed in FIG. 21 or FIG. 26 for diagnosing a pregnancy-associated disorder in a female subject pregnant with a fetus by comparing an expression level of the gene with a control value determined from one or more other female subjects, each pregnant with a healthy fetus.

In yet another aspect, the present invention provides a system comprising one or more processors configured to perform any of the above methods.

In yet another aspect, the present invention provides a computer product comprising a computer readable medium storing a plurality of instructions for controlling a processor to perform an operation of any of the methods above.

**[0005]** Methods, systems, and apparatuses are provided for diagnosing pregnancy-associated disorders, determining allelic ratios, determining maternal or fetal contributions to circulating transcripts, and/or identifying maternal or fetal markers using a sample from a pregnant female subject. In some examples, the sample is blood plasma containing a mixture of maternal- and fetal-derived RNA molecules.

**[0006]** The RNA molecules are analyzed (e.g., sequenced) to obtain a plurality of reads, and the locations of these reads in a reference sequence are identified (e.g., by sequence alignment). Informative loci are identified that are homozygous for a first allele in either the mother or fetus, and heterozygous for a first allele and a second allele in the other one of the mother and fetus. The informative loci are then filtered and reads located at (e.g., aligning to) the filtered informative loci are further analyzed. In some examples, a ratio of reads corresponding to first alleles and second alleles is calculated and compared to a cutoff to diagnose a pregnancy-associated disorder. In some examples, a portion of RNA in the sample that is of fetal origin is determined using reads located at filtered informative maternal loci. In some examples, a ratio of reads corresponding to

the first allele and the second allele is calculated for an individual filtered informative locus, and the ratio is compared to a cutoff to designate the locus as a maternal or fetal marker.

**[0007]** The present methods are not limited to prenatal diagnostics and can be applied to any biological sample that contains a mixture of RNA molecules derived from two

individuals. For example, a blood plasma sample obtained from an organ transplant recipient can be used. Transcripts expressed in the transplanted organ reflect the donor's genotype and occur at detectable levels in the receipient's blood. By reading these transcripts, informative loci within genes expressed by both the donor and recipient can be identified, and the relative expression levels of alleles from each individual can be measured. Abnormal expression levels of the allele contributed solely by the donor or recipient can be used to diagnose transplantation-associated disorders.

**[0008]** Biological samples that can be used in the present methods include blood, plasma, serum, urine, saliva and tissue samples. For example, fetal nucleic acids have been detected in the urine of pregnant women. The urine of recipients of kidney transplantation have been shown to contain cell-free nucleic acids and cells from the transplanted organ. Microchierism has been observed in many conditions. Microchimerism refers to the presence of a source of cells or nucleic acids from another person in the body, including organs and tissues of a particular individual. Microchimerism has been observed in biopsies of the thyroid, liver, spleen, skin, bone marrow, and other tissues. Microchimerism may occur as a result of previous pregnancies, or blood transfusions.

**[0009]** Also provided is use of a gene for diagnosing a pregnancy-associated disorder in a pregnant female subject. The expression level of the gene is compared with a control value determined from one or more other female subjects, each pregnant with a healthy fetus. Pregnancy-associated disorders addressed herein include pre-eclampsia, intrauterine growth restriction, invasive placentation and pre-term birth. Other pregnancy-associated disorders may be conditions that place the fetus at risk of fetal demise, such as hemolytic disease of the newborn, placental insufficiency, hydrops fetalis, fetal malformation. Yet other pregnancy-associated diseases may be conditions that result in complications during pregnancy, such as the HELLP syndrome, systemic lupus erythematosus and other immunological diseases of the mother.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0010]** FIG. 1 shows a positive relationship between the allelic ratio B/A and the relative expression levels of the respective genes in placenta and maternal blood cells.

**[0011]** FIG. 2 shows a relationship between total gene expression level and the relative expression levels of genes in placenta and maternal blood cells.

[0012] FIG. 3 is a flowchart for methods of diagnosing a pregnancy-associated disorder.

[0013] FIGS. 4 and 5 show summaries of RNA-seq read alignment results.

[0014] FIG. 6 shows data from sequencing the blood plasma RNA of two individuals. FIG. 6A shows the GC% distribution of sequenced reads from the RNA-seq libraries of M9356P (with Ribo-Zero Gold pre-treatment) and M9415P (without Ribo-Zero Gold pre-treatment). FIG. 6B shows the correlation of gene expression profiles between M9356P and M9415P.

[0015] FIG. 7 shows the RNA-SNP allelic ratios and the total plasma concentrations of different RNA transcripts in the maternal plasma plotted against the relative tissue expression levels (placenta/blood cells).

[0016] FIG. 8 shows the RNA-SNP allelic ratios and total plasma levels of different genes containing informative SNPs plotted against the relative tissue expression level (blood cells/placenta).

[0017] FIG. 9 shows RNA-SNP allelic ratios for fetal-specific SNP alleles and the expression levels of RNA transcripts containing these alleles in maternal plasma.

[0018] FIG. 10 shows RNA-SNP allelic ratios for maternal-specific SNP alleles and the expression levels of RNA transcripts containing these alleles in maternal plasma.

[0019] FIG. 11 shows data for maternal-specific SNP alleles in pre-eclamptic and control subjects. FIG. 11A shows RNA-SNP allelic ratios of case 5641 (third-trimester late-onset preeclampsia) and case 7171 (control) for informative SNPs comprising maternal-specific SNP alleles. FIG. 11B shows fold-differences of RNA-SNP allelic ratios and of plasma levels of case 5641 relative to those of 7171 (control) for informative SNPs comprising maternal-specific SNP alleles.

[0020] FIG. 12 shows data for maternal-specific SNP alleles in pre-eclamptic and control subjects. FIG. 12A shows RNA-SNP allelic ratios of case 5641 (third-trimester late-onset preeclampsia) and case 9356 (control) for informative SNPs comprising maternal-specific SNP alleles. FIG. 12B shows fold-differences of RNA-SNP allelic ratios and of plasma levels of case 5641 relative to those of 9356 for informative SNPs comprising maternal-specific SNP alleles.

[0021] FIG. 13 shows the data in FIG. 11 in tabular form.

[0022] FIG. 14 shows the data in FIG. 12 in tabular form.

[0023] FIG. 15 is a flowchart for methods of determining, in a sample from a female subject pregnant with a fetus, a portion of RNA that is of fetal origin

[0024] FIG. 16 shows results of informative SNP analysis of the maternal plasma transcriptome.

[0025] FIG. 17 shows relative fetal and maternal contributions without allele-specific expression filtering in the maternal plasma samples of late pregnancy.

[0026] FIG. 18 shows fetal and maternal contributions in the maternal plasma transcriptome. FIG. 18A shows proportions of fetal- and maternal-derived transcripts in the maternal plasma of early and late pregnancies. FIG. 18B shows allele counts and fetal-allele ratios before and after delivery.

[0027] FIG. 19 compares fractional fetal and maternal contributions at one RNA-SNP for pre-eclamptic and control pregnancy cases.

[0028] FIG. 20 is a flowchart for methods of designating a genomic locus as a maternal or fetal marker.

[0029] FIG. 21 is a list of pregnancy-associated genes.

[0030] FIG. 22 shows lower expression levels of pregnancy-associated genes in subjects post-delivery as compared with pre-delivery.

[0031] FIG. 23 shows hierarchical clustering of plasma samples using the 131 pregnancy-associated genes.

[0032] FIG. 24 shows expression levels of 131 pregnancy-associated genes in the placenta, maternal bloods cells and maternal plasma. FIG. 24A is a heatmap of gene expression ($\log_2$(transcript level)) of the placenta and maternal blood cells, as well as the pre- and post-delivery maternal plasma of the two late pregnancy cases. The 131 pregnancy-associated genes were preferentially expressed in the placenta. FIG. 24B shows that expression levels of the 131 genes in the placenta and in plasma were positively correlated ($P < 0.05$, Spearman correlation).

[0033] FIG. 25 compares fetal and maternal contributions to RNA in the plasma for genes identified with plasma- and tissue-based methods.

**[0034]** FIG. 26 lists 98 preeclampsia-associated RNAs in maternal plasma as identified by RNA-Seq. The expression levels in the plasma collected from uncomplicated pregnant women and women who have developed preeclampsia are shown.

**[0035]** FIG. 27 shows RNA-sequencing read counts on single-nucleotide polymorphism (SNP) sites in *PAPPA* RNA for case 9415. Allele-A is the shared allele and allele-G is the fetal-specific allele.

**[0036]** FIG. 28 shows RNA-sequencing read counts on single-nucleotide polymorphism (SNP) sites in *H19* RNA for case 9415.

**[0037]** FIG. 29 shows detection of maternal *H19* methylation status by methylation-sensitive restriction enzyme digestion.

**[0038]** FIG. 30 shows a block diagram of an example computer system 3000 usable with system and methods according to embodiments of the present invention.

## DETAILED DESCRIPTION

### I. INTRODUCTION

**[0039]** The presence of cell-free fetal RNA in maternal plasma was reported more than a decade ago [11]. Following this finding, many studies have since been conducted to detect circulating RNA of fetal and placental origin in maternal plasma [1,10,12]. Interestingly, the expression levels of placenta-specific transcripts in the plasma were found to be positively correlated with those in placental tissues [10], underscoring the clinical utility of plasma RNA analysis as a noninvasive tool to monitor the placental or fetal health and development. Indeed, examination of maternal circulating RNA has found clinical applications for pregnancy- or placenta-related disorders such as preeclampsia [2,13-15], intrauterine growth retardation [4] and preterm birth [8], as well as for noninvasive testing of fetal chromosomal aneuploidies [5,7,16]. Such developments highlight the potential utilities of RNA biomarkers for the molecular assessment of prenatal disorders.

**[0040]** Despite the promising outlook of plasma RNA analysis in prenatal testing, there remain a limited number of well-validated pregnancy- or placenta-related transcripts in maternal plasma to date. In this regard, examination of RNA biomarkers in the plasma had been conducted using reverse transcriptase (RT)-PCR [1,4,8,10,12,13], a sensitive method that would typically target a relatively small number of RNA species per analysis. Of note, these

studies mainly focused on genes with relatively high levels of expression in the placenta when compared with maternal blood cells, on the premise that pregnancy-associated genes are largely derived from the placenta. Such approaches have identified pregnancy-associated RNA targets that are highly expressed in the placenta, but might have missed other important targets. Furthermore, given the low concentrations and poor integrity of plasma RNA [9,15], conventional high-throughput methods such as serial analysis of gene expression and microarray analysis would not be robust for the direct examination of the plasma transcriptome.

**[0041]** The aforesaid technical limitations of plasma RNA analysis could potentially be resolved by employing massively parallel sequencing (MPS) for RNA analysis, namely RNA-sequencing (RNA-seq) [17, 18]. Given the enhanced sensitivity and wide dynamic range, RNA-seq has been employed to examine gene expression in many human tissues, including the placenta [19]. The superiority of MPS has been further demonstrated by its feasibility in direct profiling of plasma miRNAs in healthy individuals [20] and in pregnant women [21,22]. Nevertheless, the full spectrum of plasma transcriptome remains elusive, which might be due to the lower stability of long RNA species compared to short miRNAs in the plasma [23].

**[0042]** In this study, we have shown that fetal- and maternal-derived transcripts can be detected in maternal plasma and that their relative contributions can be estimated using RNA-seq, by examination of fetal- and maternal-specific single-nucleotide polymorphisms (SNPs). In addition, the allele-specific expression patterns of the placenta can be monitored in maternal plasma. We have also demonstrated that pregnancy-associated transcripts could be identified by direct examination of the maternal plasma before and after delivery.

## II. METHOD OF DIAGNOSING A PREGNANCY-ASSOCIATED DISORDER USING SNP ALLELIC RATIOS

**[0043]** The analysis of the gene expression profile is useful for the detection of the disease status of an individual. Embodiments of the present invention include methods to analyze the expression profile of an individual through the analysis of a mixture of RNA molecules from two different individuals. The methods make use of the relative abundances of alleles that are specific to one individual and the alleles shared between the two individuals. Based on the relative abundances of the shared and individual-specific alleles, the gene expression profiles of the two individuals can be determined. One possible application of the present methods is the analysis of the gene expression profile of a fetus by analysing the RNA in a maternal

plasma sample which contains RNA from both the fetus and the pregnant woman. Another application is the analysis of the gene expression profile of donor-derived RNA in a plasma sample collected from a transplantation recipient that contains RNA from both the donor and recipient.

**[0044]** In a mixture containing RNA from two individuals, the expression profile of each individual cannot be determined based on the analysis of the total amount of the different RNA transcripts in the mixture because it is difficult to determine the relative contributions of each individual to the total RNA. The relative abundance of different RNA transcripts can be useful for monitoring of the well-being of an individual or to detect a disease state.

**[0045]** In this method, the genotype of the two individuals can first be determined either by direct genotyping of the individuals or by family analysis. For example, if the parents have the genotype of AA and TT, then the genotype of the fetus would be AT.

**[0046]** The following hypothetical example illustrates the principle of this method. For each of the gene of interest, there is a SNP within the coding region so that polymorphism can be observed in the RNA transcripts of the different genes. Assume that the genotypes of the fetus and the pregnant woman are AB and AA, respectively. Thus, the B allele would be specific for the fetus and the A allele would be shared between the fetus and the mother. The relative expression levels of the different genes in the placenta and the maternal blood cells are as shown in **TABLE 1**.

**[0047]** In this example, we assume that the placenta and maternal blood cells would contribution a relative equal proportion of 2% of each of their RNA transcripts to the maternal plasma. In other words, if the placenta expression level of a gene is 10000, it would contribute 200 RNA transcripts of that gene to the maternal plasma. As shown in **FIG. 1**, the allelic ratio B/A shows a good positive relationship with the relative expression levels of the respective genes in placenta and maternal blood cells. In contrast, the total gene expression level is affected by the fluctuation of the expression of the maternal blood cells and hence, correlated less well with the expression of the placenta (**FIG. 2**).

Table 1.

| Gene | Maternal genotype | Fetal genotype | Relative expression level | | Relative expression (placenta /maternal blood cells) | Maternal plasma | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | in maternal blood cells | in the placenta | | Relative abundance of maternal to fetal RNA | Total RNA conc | Conc of RNA | | B/A ratio |
| | | | | | | | | with the A allele | with the B allele | |
| 1 | AA | AB | 8000 | 20000 | 2.5 | 1:1 | 560 | 360 | 200 | 0.56 |
| 2 | AA | AB | 6000 | 10000 | 1.67 | 1:1 | 320 | 220 | 100 | 0.45 |
| 3 | AA | AB | 10000 | 5000 | 0.5 | 1:1 | 300 | 250 | 50 | 0.2 |
| 4 | AA | AB | 14000 | 1000 | 0.071 | 1:1 | 300 | 290 | 10 | 0.03 |

**[0048]** Another advantage of using the allelic ratio analysis is that the expression level of a particular gene in the placenta is normalized to the expression level of the maternal blood cells. As the expression level can vary greatly across different genes and different samples, this normalization would make the comparison across different genes more robust and avoid the necessity of comparing to a reference gene, for example, a house keeping gene. In other words, in conventional gene expression analysis, level of expression of a gene in a sample is either measured with reference to a house-keeping gene or as an amount to the total RNA of the sample. To identify aberrant gene expression, conventionally, one would then compare those relative values of the test sample to the control samples.

**[0049]** Here we propose a new approach whereby we use the relative amounts contributed by the fetus normalized to the mother's contribution of the same gene as a means to determine the gene expression profile. In pathological states, for example preeclampsia, preterm labor, maternal diseases such as systemic erythematosus, whereby the expression of the gene in the placenta or maternal organs are altered, the fetal to maternal ratio of that gene would be altered when compared with pregnancies without such pathological conditions. This approach can be practised using the fetal-specific SNP allele relative to the shared allele among RNA transcripts in maternal plasma.

**[0050]** The approach can also be practised using the maternal-specific SNP allele relative to the shared allele among RNA transcripts in maternal plasma. Pathological conditions could be identified when one or more of such allelic ratios for one or more RNA gene transcripts are altered when compared to that expected for a non-pathological state. The pattern of such allelic ratios across a gene locus or multiple gene loci could be used to identify pathological states. Non-pathological states could be represented by the allelic ratios in normal pregnancies, in samples obtained before or after testing when the pregnancy is no longer

affected by the pathological state or from existing data previously obtained from normal pregnancies, i.e. a previously derived reference range.

**[0051]** Pregnancy-associated disorders that can be diagnosed using the present methods include any disorders characterized by abnormal relative expression levels of genes in maternal and fetal tissue. These disorders include, but are not limited to, preeclampsia, intrauterine growth restriction, invasive placentation, pre-term birth, hemolytic disease of the newborn, placental insufficiency, hydrops fetalis, fetal malformation, HELLP syndrome, systemic lupus erythematosus, and other immunoligcal diseases of the mother. The methods distinguish between RNA molecules contributed by the mother and fetus in a sample that contains a mixture of such molecules. The methods can thus identify changes in the contribution from one individual (i.e., the mother or fetus) to the mixture at a particular locus or for a particular gene, even if the contribution from the other individual does not change or moves in the opposite direction. Such changes cannot be easily detected when measuring the overall expression level of the gene without regard to the tissue or individual of origin. Pregnancy-associated disorders, as discussed herein, are not characterized by chromosomal abnormalities in the fetus, for example aneuploidy.

**[0052]** The present methods can also be performed without prior genotyping information of the placenta. For example, the genotype of the pregnant woman can be determined by direct genotyping of her blood cells. Then, the RNA transcripts of the plasma samples can be analyzed, for example but not limited to massively parallel sequencing. RNA transcripts showing two different alleles can be identified. In this set of transcripts, if the mother is homozygous, it would indicate that the fetus is heterozygous for the fetal-specific allele and the maternal allele. In this scenario, the RNA allelic ratio analysis can be performed without prior genotype information of the fetus (or the placenta). Additionally, instances where the mother is heterozygous but the fetus is homozygous can be identified by a deviation from a 1:1 ratio. Further details of such techniques are described in U.S. Patent 8,467,976.

### *A. Diagnostic method*

**[0053]** A method 300 for diagnosing a pregnancy-associated disorder according to some embodiments is shown in **FIG. 3**. The method uses a sample from a female subject pregnant with a fetus. The sample can be of maternal blood plasma and contains a mixture of maternal- and fetal-derived RNA molecules. The sample can be obtained as desired, from a female subject at any stage of pregnancy. For example, first-, second- and third-trimester

pregnant women with singleton pregnancies can serve as subjects. First- and second-trimester pregnant women can be categorized as "early pregnancy cases" and third-trimester pregnant women can be categorized as "late pregnancy cases". In some embodiments, samples can also be collected after delivery of the fetus, or from the same subject before and after delivery, and samples from non-pregnant females can serve as controls. Samples can be obtained any time after delivery (for example, 24 hours).

**[0054]** The sample can be obtained from peripheral blood. A maternal blood sample can be processed as desired, for example by centrifguing to separate blood cells from plasma. Stablizers can be added the blood sample or portions thereof, and the sample can be stored before use. In some embodiments, samples of fetal tissue are also obtained, for example chorionic villi, amniotic fluid, or placental tissue. The fetal tissue can be used to determine fetal genotypes, as is discussed below. Fetal tissue can be obtained before or after delivery.

**[0055]** Once a sample is obtained, RNA or DNA can then be extracted from the maternal blood cells and plasma, as well as from any fetal tissue such as a sample of the placenta. Placental and blood cell RNA samples can be pre-treated, such as with the Ribo-Zero Gold Kit (Epicentre), to remove ribosomal RNA (rRNA) prior to sequencing library preparation.

**[0056]** In step 301 of the method, a plurality of reads is received. The reads are obtained from an analysis of RNA molecules obtained from the sample. In various embodiments, the reads can be obtained by sequencing, digital PCR, RT-PCR, and mass spectrometry. Although the discussion focuses on sequencing, aspects of the description are also applicable to other techniques for obtaining the reads. For example, digital PCR (including microfluidic and droplet PCR) can provide knowledge of different alleles at a particular locus using probes (including primers) directed to each allele. The probes can bear labels to make them distinguishable from each other, for example by exhibiting different colors. In the same experiment or in a different experiment (for example, on a separate slide or chip), probes with different labels can be directed to different loci. Detection of the labels reflects the presence and/or quantity of the RNA or cDNA molecules being amplified and corresponds to reads of the RNA or cDNA molecules, where such reads provide information about a sequence of the RNA or cDNA molecules at the loci corresponding to the probes. The description regarding "sequence reads" equally applies to reads obtained via any suitable technique, including non-sequencing techniques.

[0057] Sequencing can be performed as desired, using any available technology. Examples of nucleic acid sequencing technologies and methods include massively parallel sequencing, next-generation sequencing, whole genome sequencing, exome sequencing, sequencing with or without target enrichment, sequencing by synthesis (e.g., sequencing by amplification, clonal amplification, bridge amplification, sequencing with reversible terminators), sequencing by ligation, sequencing by hybridization (e.g., microarray sequencing), single-molecule sequencing, real-time sequencing, nanopore sequencing, pyrosequencing, semiconductor sequencing, sequencing by mass spectroscopy, shotgun sequencing, and Sanger sequencing. In some embodiments, sequencing is performed using massively parallel techniques on a cDNA library prepared from RNA in the sample. cDNA libraries can be synthesized as desired, for example using the mRNA-Seq Sample Preparation Kit (Illumina) following the manufacturer's instructions or with slight modifications. In some embodiments, 5-fold diluted Klenow DNA polymerase is used for the end-repair step of plasma cDNA. The QIAquick PCR Purification Kit and the QIAquick MinElute Kit (Qiagen) can be used for purifying end-repaired and adenylated products, respectively. In some embodiments, 10-fold diluted paired-end adapters are used for plasma cDNA sample, or two rounds of purification are performed for adapter-ligated products using the AMPure XP beads (Agencourt). A cDNA library can be sequenced for 75 bp in a paired-end format on a HiSeq 2000 instrument (Illumina), or using other formats or instruments.

[0058] In step 302, locations of the reads in a reference sequence are determined. For PCR-based techniques, a location can be determined, for example, by matching the color of a detected label on a probe or primer to the sequence for which the probe or primer is specific. For sequencing techniques, the location can be determined by aligning the sequence read to the reference sequence. Sequence alignment can be performed by a computer system. Any bioinformatics pipeline can be used for raw data preprocessing (such as removal of highly duplicated reads and reads of low quality), data alignment and/or data normalisation. The RNA transcript levels in the maternal blood cells, placental tissues and maternal plasma can be calculated as fragments per kilobase per millionth exonic reads (FPKM). For determining allelic ratios, data preprocessing and alignment can be performed but data normalisation is not required. Any reference sequence (for example, the hg19 reference human genome) and any algorithm can be used for alignment.

[0059] In one example of carrying out steps 301 and 302, after removal of duplicated reads and rRNA reads, an average of 3 million analyzable reads were obtained for each non-

pregnant female plasma sample; an average of 12 million analyzable reads were obtained for each plasma sample of pregnant women. For tissue RNA-seq, an average of 173 million and 41 million analyzable reads per sample were obtained for placenta and blood cells, respectively. The RNA-seq alignment statistics are summarized in **FIGS. 4 and 5**. The GC content of sequenced reads was also examined in all samples (**FIG. 6A** and **TABLE 2**).

**Table 2. Proportion of the GC-rich sequenced reads in RNA-seq libraries.**

| Sample type | Sample | Treatment | GC-rich reads (% of raw reads)[a] |
|---|---|---|---|
| Plasma | M9356P | rRNA-depletion | 49.28% |
| | M9415P | None | 0.00% |
| Placenta | N9356 | rRNA-depletion | 0.15% |
| Blood cells | M9356W | rRNA-depletion | 0.30% |

*a*. Raw reads were aligned against the *de novo* assembled GC-rich sequence by BLAST (NCBI).

**[0060]** In step 303, one or more informative loci are identified. This step can determine or make inferences about the genotypes of the female subject and fetus at loci within the genome, and compare these genotypes. A locus is considered informative if it is homozygous in a first entity for a corresponding first allele (e.g., AA), and is heterozygous in a second entity for the corresponding first allele and a corresponding second allele (e.g., AB). The first entity can be the pregnant female subject or the fetus, and the second entity is the other one of the pregnant female subject and the fetus. In other words, one individual (either the mother or fetus) is homozygous, and the other individual is heterozygous, for each informative locus.

**[0061]** Informative loci can be further classified according to which individual is heterozygous and the sole contributor of one allele. A locus is considered an informative maternal locus, or equivalently a maternal-specific locus, if the fetus is homozygous and the pregnant female subject is heterozygous. A locus is considered an informative fetal locus, or equivalently a fetal-specific locus, if the pregnant female subject is homozygous and the fetus is heterozygous. In step 303, the informative loci identified are all either maternal-specific or fetal-specific, because the same individual serves as the second entity for all of these loci.

**[0062]** An informative locus can represent a single-nucleotide polymorphism or "SNP", where the first allele and second allele differ in the identity of a single nucleotide. An informative locus can also represent a short insertion or deletion, where one or more nucleotides are inserted or deleted in one allele as compared with the other allele.

[0063] In some embodiments, the genotype of the pregnant female subject at one or more informative loci, or at each informative locus, is determined by sequencing genomic DNA obtained from maternal tissue. The maternal tissue can be maternal blood cells or any other kind of tissue. In some embodiments, the genotype of the fetus at one or more informative loci, or at each informative locus, is determined by sequencing genomic DNA obtained from fetal tissue, such as the placenta, chorionic villi, amniotic fluid. If a less invasive method is preferred, fetal DNA can also be obtained for sequencing from a maternal blood sample. The same sample used to obtain the RNA molecules for sequencing can be the source of such fetal DNA, or a different sample can be used. It will be appreciated that informative fetal loci can be identified without directly genotyping the fetus. For example, if the pregnant female subject is determined to be homozygous for a first allele at one locus, and the RNA sequencing indicate the presence of a second allele in the mixture of maternal and fetal RNA, then the fetus can be assumed to be heterozygous at this locus.

[0064] In some embodiments, the mother and fetus are both genotyped using massively parallel methods to identify informative loci. Sequencing can be performed on exome-enriched maternal blood cell and placental genomic DNA samples. Genomic DNA can be extracted from the placental tissues and maternal blood cells using the QIAamp DNA Kit and the QIAamp Blood Kit (both from Qiagen), respectively, following the manufacturer's instructions. Exome enrichment and sequencing library preparation can be performed using the TruSeq Exome Enrichment Kit (Illumina) following the manufacturer's protocol. The libraries can be sequenced for 75 bp in a PE format on a HiSeq 2000 instrument (Illumina).

[0065] In step 304, the one or more informative loci are filtered. As used herein, "filtering" means selecting certain informative loci, out of those identified in step 303, for further analysis. A "filtered" informative locus is a locus so selected. The selection can be made on the basis of one or more criteria. One such criterion is where the informative loci are located in the genome or on a reference sequence. In some embodiments, only loci located within an exon or expressed region of the reference sequence are further analyzed.

[0066] Another criterion for filtering is the number of sequence reads, out of all of those received in step 301 and aligned to the reference sequence in step 302, that align to the locus and contain each allele. In some embodiments, a filtered informative locus must have associated with it at least a first predetermined number of sequence reads containing the first allele, and/or a second predetermined number of sequence reads containing the second allele.

The predetermined numbers can be 1, 2 or more, and can correspond to a desired read quality or sequencing depth. As a result of filtering, one or more filtered informative loci are identified.

**[0067]** In some embodiments, only informative loci representing SNPs are considered for filtering. In one example of the method, approximately one million SNPs in the NCBI dbSNP Build 135 database, all located in exons, were examined, and informative SNPs were identified out of these million using maternal and fetal genotypes. For illustration of informative SNPs, "A" was assigned as the shared allele and "B" as the maternal- or fetal-specific allele. Informative SNPs included those with a maternal-specific SNP allele, i.e. "AA" in the fetus and "AB" in the mother on a given locus, and those with a fetal-specific SNP allele, i.e. "AA" in the mother and "AB" in the fetus, on a given locus. The genotypes were called using an in-house bioinformatics pipeline. To perform filtering, only informative SNPs in which both the "A" allele and the "B" allele showed at least one read count, respectively, were included for analysis.

**[0068]** In steps 305 and 306, the sequence reads aligning to each filtered informative locus are counted and sorted depending on which allele they contain. For each filtered informative locus, two numbers are determined: a first number, which is the number of sequence reads aligning to the locus and containing the corresponding first allele (i.e., the shared allele); and a second number, which is the number of sequence reads aligning to the locus and containing the corresponding second allele (i.e., the maternal- or fetal-specific allele). The first number and second number can be computed as desired. The first number plus the second number for each locus provides a total number of sequence reads aligning to the locus. A ratio of the first number and the second number for a particular filtered informative locus can be referred to as an "allelic ratio".

**[0069]** In steps 307 and 308, the numbers of the first alleles and second alleles are summed across the filtered informative loci. In step 307, a first sum of the first numbers is calculated. The first sum represents the total number of sequence reads containing the corresponding first alleles (i.e., the shared alleles) for the filtered informative loci. In step 308, a second sum of the second numbers is calculated. The second sum represents the total number of sequence reads containing the corresponding second alleles (i.e., the maternal- or fetal-specific alleles) for the filtered informative loci. The sums can be computed as desired, using all filtered informative loci identified in step 304 or a subset thereof. In some embodiments, the sums

are weighted. For example, the contributions to the sums of loci from a certain chromosome can be scaled up or down by multiplying the first numbers and second numbers for those loci by a scalar.

**[0070]** In step 309, a ratio of the first sum and the second sum is calculated. This ratio represents the relative numbers of sequence reads for the first alleles and second alleles, aggregated across filtered informative loci. In some embodiments, the ratio is simply calculated as the first sum divided by the second sum. Such a calculation yields the number of sequence reads for shared alleles as a multiple of sequence reads for maternal- or fetal-specific alleles. Alternatively, the ratio can be calculated as the second sum divided by the sum of the first sum and the second sum. Here the ratio provides the number of sequence reads for maternal- or fetal-specific alleles as a fraction of all sequence reads. Other methods of calculating the ratio will be apparent to the skilled artisan. The first sum and second sum, from which the ratio is calculated, can serve as proxies for the amounts of RNA (i.e., transcripts) present in the sample that originate from shared and specific alleles, for the filtered informative loci.

**[0071]** In step 310, the ratio calculated in step 309 is compared to a cutoff value to determine whether the fetus, mother, or pregnancy has a pregnancy-associated disorder. In various embodiments, a cutoff value can be determined from one or more samples obtained from pregnant female subjects without the pregnancy-associated disorder (i.e., control subjects) and/or determined from one or more samples obtained from pregnant female subjects with the pregnancy-associated disorder. In some embodiments, the cutoff is determined by performing the same method as described above on samples from control subjects, and examining the same or an overlapping set of filtered informative loci. A cutoff can be set at a value between expected values for pregnancies without a disorder and expected values for pregnancies with a disorder. The cutoff can be based on a statistical difference from a normal value.

**[0072]** In some embodiments, the ratio calculated for the pregnant female subject (as in step 309) uses filtered informative loci in a certain set of genes, and a ratio (i.e. the cutoff) is calculated for the control subjects using filtered informative loci in some or all the same genes. If the ratio calculated for the pregnant female subject is based on maternal-specific alleles, then the cutoff calculated for the control subjects can also be based on maternal-specific alleles. Similarly, the ratio and cutoff can both be based on fetal-specific alleles.

**[0073]** The comparison between the ratio and cutoff can be performed as desired. For example, a disorder can be diagnosed if the ratio exceeds the cutoff or falls below the cutoff, by any amount or by a certain margin. The comparison can involve calculating a difference or calculating another ratio between the ratio for the pregnant female subject and the cutoff value. In some embodiments, the comparison involves evaluating whether the relative expression levels of shared and unshared alleles for certain genes significantly differ between the pregnant female subject and the control subjects.

**[0074]** If desired, the method can also be used to estimate a portion of RNA in the sample that is of maternal or fetal origin. The estimate is made by multiplying the ratio calculated in step 309 by a scalar. The scalar represents a total expression level at the filtered informative loci relative to expression of the second alleles in the heterozygous individual.

**[0075]** The following example illustrates application of the scalar. If the second alleles counted in the method are fetal-specific, then the ratio calculated in step 309 can provide the number of sequence reads for fetal-specific alleles as a fraction of all sequence reads for the filtered informative loci. Of the sequence reads containing the shared allele, some are contributed by the mother and some are contributed by the fetus. If the relative expression levels of the fetal-specific and shared alleles in the fetus are known, or can be estimated, then the ratio can be scaled to obtain an estimate of the fractional fetal contribution to all sequence reads for the filtered informative loci.

**[0076]** In some embodiments, the fetal and shared alleles are expressed symmetrically at most loci and the scalar is assumed to be about two In other embodiments, the scalar departs from two and takes into account asymmetric gene expression [24]. The fractional maternal contribution to sequence reads is then one minus the fractional fetal contribution. Similar calculations can be made when the second alleles at the filtered informative loci are maternal-specific.

### *B. Examples:*

#### 1. Correlation of allelic ratios and plasma concentrations

**[0077]** The expression profile of a maternal plasma sample (gestational age 37 2/7) was analyzed and compared with the corresponding placenta and maternal blood cells for a pregnant woman. Massively parallel sequencing of RNA was performed for each of these samples using the Illumina HiSeq 2000 instrument. The placenta and the maternal blood cells

were genotyped using exome sequencing by massively parallel sequencing. We examined approximately one million SNPs in the NCBI dbSNP Build 135 database that were located in the exons, and sorted for informative SNPs where the mother is homozygous (genotype AA) and the fetus is heterozygous (genotype AB). Thus, the A allele would be the allele shared between the mother and the fetus and the B allele is fetal-specific.

**[0078]** The RNA-SNP allelic ratios and the total plasma concentrations of different RNA transcripts in the maternal plasma are plotted against the relative tissue expression levels (placenta/blood cells) (**FIG. 7**). We can observe that the RNA-SNP allelic ratio in the plasma correlates well with the relative tissue expression level (Spearman R = 0.9731679, $P$ = 1.126e-09) However, the total plasma level did not correlate with the tissue expression (Spearman R = -0.7285714, $P$ = 0.002927).

**[0079]** In addition to the analysis of the fetal expression, this method can also be used for the profiling of the maternal expression. In this scenario, informative SNPs that the mother is heterozygous (genotype AB) and the fetus is homozygous (genotype AA) can be used. The RNA-SNP allelic ratio can then be calculated as the maternal-specific allelic count divided by the count of the shared allele. In **FIG. 8**, the RNA-SNP allelic ratios and total plasma levels of different genes containing informative SNPs are plotted against the relative tissue expression level (blood cells/placenta). We can observed a good positive relationship between the tissue expression level with the RNA-SNP allelic ratio in the maternal plasma (Spearman R=0.9386, $P$<2.2e-16) but not the total plasma transcript levels (Spearman R=0.0574, $P$=0.7431).

**[0080]** The RNA transcripts, their RNA-SNP allelic ratios and their plasma levels are tabulated in **FIG. 9 and FIG. 10.**

**2. Comparison of maternal plasma RNA-SNP allelic ratio for preeclampsia and control pregnancy cases**

**[0081]** For the cases of two pregnant female subjects, an average number of 117,901,334 raw fragments were obtained (**TABLE 3A**). The RNA-SNP allelic ratio of a subset of circulating transcripts bearing maternal-specific SNP alleles was compared between a third-trimester late-onset preeclampsia (PET) case, 5641, and its gestational age-matched control case, 7171. As shown in **FIG. 11A**, the RNA-SNP allelic ratios of these transcripts portrayed a different profile in the PET and the control cases. Importantly, the RNA-SNP allelic ratio profile is distinct from the plasma transcript level profile, as shown in **FIG. 11B**. This

suggests that the maternal plasma RNA-SNP allelic ratio analysis could be utilized as another metric to better distinguish PET cases from a normal pregnancy cases. The number of informative transcripts could be further increased with an increased number of informative SNPs, exemplified by the comparison between the PET case, 5641, and another normal pregnancy control case, 9356 (**FIGS. 12A** and **12B**). The RNA-SNP allelic ratios and the plasma level of the RNA transcripts shown in **FIG. 11** and **FIG. 12** are tabulated in **FIG. 13** and **FIG. 14**, respectively.

**Table 3A: Summary of RNA-seq read alignment results.**

| Sample type | Sample | Raw reads | Preprocessed[a] reads | (%)[b] | Mappable reads[c] | (%)[d] |
|---|---|---|---|---|---|---|
| Maternal plasma | 5641 (Preeclampsia) | 119,902,178 | 30,106,800 | 25.11% | 4,001,376 | 13.29% |
| Maternal plasma | 7171 (Control) | 115,900,490 | 29,160,920 | 25.16% | 1,397,345 | 4.79% |

*a.* Retained reads after removal of highly repetitive reads.
*b.*% of the raw reads.
*c.* See Supplemental Data Table 2B for detailed breakdown.
*d.* % of the preprocessed reads.

**Table 3B: Summary of RNA-seq read alignment results**

| Sample type | Sample | Filtered reads[a] | Analyzable reads[b] | Exon (%) | Intron (%) | Intergenic region (%)[c] |
|---|---|---|---|---|---|---|
| Maternal plasma | 5641 (Preeclampsia) | 2,579,163 | 1,422,213 | 44.87% | 12.32% | 42.80% |
| Maternal plasma | 7171 (Control) | 790,735 | 606,610 | 37.69% | 16.09% | 46.22% |

*a.* Filtered reads are mostly the nuclear and mitochondrial rRNAs and tRNAs.
*b.* Analyzable reads = total mappable reads – filtered reads.
*c.* % of reads that aligned to regions outside exons and introns of the reference genes.

## III. METHOD OF DETERMINING FETAL OR MATERNAL RNA CONTRIBUTION

### *A. Method*

**[0082]** A method 1500 is shown in **FIG. 15** for determining, in a sample from a female subject pregnant with a fetus, a portion of RNA that is of fetal origin. The sample can be obtained and prepared as discussed above for method 300.

**[0083]** In step 1501, a plurality of sequence reads are received. In step 1502, the sequence reads are aligned to a reference sequence. These steps can be performed as described above for steps 301 and 302.

**[0084]** In step 1503, one or more informative maternal loci are identified. Each informative maternal locus is homozygous in the fetus for a corresponding first allele and heterozygous in the pregnant female subject for the corresponding first allele and a corresponding second allele. The first allele is thus shared between the mother and fetus, and the second allele is maternal-specific. The informative maternal loci can be identified as described above for step 303.

**[0085]** In step 1504, the informative maternal loci are filtered, as in step 304, and one or more filtered informative maternal loci are identified. Sequence reads are then manipulated on the level of individual filtered informative maternal loci in steps 1505-1511. In steps 1505 and 1506, a first number and a second number are determined for each filtered informative maternal locus. The first number is determined in step 1505 as the number of sequence reads aligning to the locus and containing the corresponding first allele. The second number is determined in step 1506 as the number of sequence reads aligning to the locus and containing the corresponding second allele. Steps 1505 and 1506 can be performed similarly to steps 305 and 306, discussed above.

**[0086]** In step 1507, a sum of the first number and the second number is calculated for each informative maternal locus. The sum can be equivalent to the total number of sequence reads aligning to the locus. In step 1508, a maternal ratio is calculated by dividing the second number by the sum. The maternal ratio provides the fraction of all sequence reads aligning to the particular locus that contain the second (maternal-specific) allele.

**[0087]** In step 1509, a scalar is determined. The scalar represents a total expression level at the filtered informative maternal locus relative to expression of the corresponding second allele in the pregnant female subject. In some embodiments, where expression of the two alleles in the pregnant female subject is known or assumed to be symmetrical, the scalar is about two. In other embodiments, the scalar can depart from two and take into account asymmetrical gene expression. As most loci express symmetrically, the assumption of symmetry is valid.

**[0088]** In step 1510, the maternal ratio is multiplied by the scalar to obtain a maternal contribution. The maternal contribution represents, at the filtered informative maternal locus,

the fraction of sequence reads (or, by extension, the fraction of RNA in the sample) that is contributed by the mother. One minus the maternal contribution is the fetal contribution, or the fraction of sequence reads contributed by the fetus. In step 1511, the fetal contribution is calculated for the filtered informative maternal locus.

**[0089]** In step 1512, a portion of RNA in the sample that is of fetal origin is determined. The portion is the average of the fetal contributions for the filtered informative maternal loci. In some embodiments, this average is weighted, for example by the sums calculated for the filtered informative maternal loci. By calculating a weighted average, the portion determined in step 1512 can reflect the relative expression levels of different loci or genes in the sample.

**[0090]** The portion of RNA in the sample that is of fetal origin can be calculated for one filtered informative maternal locus, many such loci, or all such loci for which sequencing data are obtained. It will be appreciated that this portion reflects the circulating transcriptome in the particular pregnant female subject at the time the sample was acquired. If samples are obtained from the subject at different times over the course of the pregnancy, the set of filtered informative maternal loci identified in method 1500 may vary from one sample to the next, as may the portions of fetal-originating RNA determined at these loci. The portion of RNA in a sample that is of fetal origin can also vary from one subject to the next, even when controlling for gestational stage or other factors, and can differ between a subject having a pregnancy-associated disorder and a healthy subject. Accordingly, the portion can be compared to a cutoff-value to diagnose such a disorder.

**[0091]** In some embodiments, the cutoff value is determined by performing method 1500 using samples from one or more healthy subjects. In some embodiments, a diagnosis is made if the portion exceeds the cutoff or falls below the cutoff, by any amount or by a certain margin. The comparison can involve calculating a difference or calculating a ratio between the portion of fetal-originating RNA in the pregnant female subject and the cutoff value. In some embodiments, the comparison involves evaluating whether the portion of fetal-originating RNA in the pregnant female subject differs significantly from what is seen in healthy pregnant women at similar gestational periods.

**[0092]** The portion of RNA in the sample that is of fetal origin can also be determined by examining informative fetal loci. At each informative fetal locus, this portion can be estimated by first calculating a fetal ratio, which is the number of sequence reads containing the second (fetal-specific) allele divided by the total number of sequence reads. The fetal

contribution is then the fetal ratio multiplied by a scalar representing the relative expression levels of the two alleles in the fetus. Accordingly, a variation of method 1500 can be performed by identifying informative fetal loci, filtering these loci, calculating a fetal contribution for each filtered locus, and averaging the fetal contributions for the filtered loci. If desired, the fetal-originating portion of RNA determined by method 1500 can be augmented by averaging the fetal contributions calculated for filtered informative maternal loci with fetal contributions calculated for filtered informative fetal loci. The averages calculated here can be weighted, to reflect variation in the number of sequence reads at different loci, to give more weight to maternal- or fetal-specific loci, or otherwise as desired.

*B. Examples*

**1. Identification and estimation of fetal- and maternal-derived transcript in maternal plasma**

**[0093]** Informative genes, defined as genes with at least one informative SNP, were first identified based on the genotyping data. In the two early pregnancy cases, a total number of 6,714 and 6,753 informative genes were available to examine the relative proportions of fetal and maternal contributions, respectively (**FIGS. 16 and 17**). In the two late pregnancy cases, a total number of 7,788 and 7,761 informative genes were available to examine the relative proportions of fetal and maternal contributions, respectively. To measure the relative proportion of fetal contribution in the maternal plasma, we sorted for RNA transcripts where the fetal-specific alleles were covered by at least one RNA-seq read in the maternal plasma samples. The relative proportions of such fetal-derived transcripts were 3.70% and 11.28% during early and late gestations, respectively. Using a similar approach, the relative proportions of maternal contribution in the circulation, examined using the maternal-specific SNP alleles, was estimated to be 76.90% and 78.32% during early and late gestations, respectively (**FIG. 18A**).

**2. Comparison of RNA-SNP fractional fetal and maternal contribution for PET and control pregnancy cases**

**[0094]** We have examined the fractional fetal and maternal contribution for GNAS transcript, which was detected in both the PET case, 5641, and its gestational age-matched control case, 7171. In the GNAS transcript, there was an informative SNP site comprising fetal-specific SNP allele at locus rs7121 in case 5641; on the same SNP site in case 7171, there was a maternal-specific SNP allele. The fractional fetal and maternal contributions for

this SNP site in case 5641 were calculated to be 0.09 and 0.91, respectively. On the other hand, the fractional fetal and maternal contributions for the same SNP site in case 7171 were 0.08 and 0.92, respectively (**FIG. 19**). Compared to the control case 7171, there was a 12.5% increase in the fractional fetal contribution and a 1.09% decrease in the fractional maternal contribution in this transcript. Interestingly, a 21% increase in FPKM was detected for this transcript in case 5641 compared to 7171.

## IV. METHOD OF DESIGNATING A GENOMIC LOCUS AS A MATERNAL OR FETAL MARKER

### *A. Method*

**[0095]** A method 2000 of designating a genomic locus as a maternal or fetal marker is shown in **FIG. 20**. The method involves analyzing a sample from a female subject pregnant with a fetus. The sample can be of maternal blood plasma and contains a mixture of maternal- and fetal-derived RNA molecules. The sample can be obtained and prepared as discussed above for method 300.

**[0096]** In step 2001, a plurality of sequence reads are received. In step 2002, the sequence reads are aligned to a reference sequence. These steps can be performed as described above for steps 301 and 302.

**[0097]** In step 2003, one or more informative loci are identified. Each locus is homozygous in a first entity for a corresponding first allele and heterozygous in a second entity for the corresponding first allele and a corresponding second allele. The first entity is the pregnant female subject or the fetus, and the second entity is the other one of the pregnant female subject and the fetus. The informative loci can be identified as described above for step 303.

**[0098]** In step 2004, the informative loci are filtered, as in step 304, and one or more filtered informative loci are identified. In some embodiments, only informative loci representing SNPs are considered for filtering. In one example of the method, filtered informative loci included informative SNPs with at least one sequence read for the "A" allele and one sequence read for the "B" allele in the maternal plasma.

**[0099]** Sequence reads are then manipulated on the level of individual filtered informative loci in steps 2005-2008. In steps 2005 and 2006, a first number and a second number are determined for each filtered informative locus. The first number is determined in step 2005

as the number of sequence reads aligning to the locus and containing the corresponding first allele. The second number is determined in step 2006 as the number of sequence reads aligning to the locus and containing the corresponding second allele. Steps 2005 and 2006 can be performed similarly to steps 305 and 306, discussed above.

**[0100]** In step 2007, a ratio of the first number and the second number is calculated for each filtered informative locus. This ratio represents the relative abundances of sequence reads (and, by extension, transcripts in the sample) containing the first allele and the second allele. In some embodiments, the ratio is simply calculated as the first number divided by the second number. Such a calculation yields the number of sequence reads for the shared allele as a multiple of sequence reads for the maternal- or fetal-specific allele. This ratio, or its reciprocal, can be considered an allelic ratio, and for SNP loci, an RNA-SNP allelic ratio.

**[0101]** In some embodiments, for informative SNPs that include maternal-specific SNP alleles, the RNA-SNP allelic ratio for each SNP is calculated as maternal-specific allele:shared allele. On the other hand, for informative SNPs that include fetal-specific SNP alleles, the RNA-SNP allelic ratio can be calculated as fetal-specific allele:shared allele. Unlike gene expression analysis, in which data normalisation is a pre-requisite, RNA-SNP allelic ratio analysis does not require data normalisation and less bias is introduced. For transcripts that contain more than one informative SNP, an RNA-SNP allelic ratio can be calculated for each informative SNP site and an average RNA-SNP allelic ratio per transcript can be computed.

**[0102]** In step 2007, the ratio can alternatively be calculated as the second number divided by the sum of the first number and the second number. Here the ratio provides the number of sequence reads for the maternal- or fetal-specific allele as a fraction of all sequence reads at the locus. In terms of "A" and "B" alleles, the ratio can be expressed as

$$\text{B-allele ratio} = \text{B-allele count} / (\text{A-allele count} + \text{B-allele count})$$

Theoretically, for a plasma transcript that is contributed solely by the fetus or by the mother, the B-allele ratio should be 0.5, assuming no allele-specific expression. Other methods of calculating the ratio will be apparent to the skilled artisan.

**[0103]** In step 2008, the filtered informative locus is designated as a marker when the ratio exceeds a cutoff. In some embodiments, the cutoff is from about 0.2 to about 0.5. In some embodiments the cutoff is 0.4. In one example of the method, a B-allele ratio cutoff was

defined as ≥ 0.4 for an RNA transcript with high fetal or maternal contribution. This cutoff took into consideration the Poisson distribution and random sampling of RNA-seq reads. In some embodiments, the contribution of the "B" allele is said to be high when the ratio exceeds the cutoff.

**[0104]** A high allelic ratio at a particular informative locus can indicate that (i) the second or "B" allele is more highly expressed in the heterozygous individual than the "A" allele (i.e., the allele is expressed asymmetrically), (ii) a greater share of all RNA aligning to the locus in the maternal plasma is contributed by the heterozygous individual, or both (i) and (ii). A high allelic ratio can also indicate a pregnancy-associated disorder if the gene associated with the informative locus is being pathologically overexpressed. Accordingly, some embodiments of the method also include diagnosing a pregnancy-associated disorder based upon whether a filtered informative locus is designated as a marker for the second entity (i.e., the heterozygous individual). In making such a diagnosis, the cutoff used for comparison to the allelic ratio can be based on transcript levels in plasma samples obtained from healthy pregnant subjects.

**[0105]** In some embodiments, method 2000 can also be used to estimate a portion of RNA in the sample, at a particular filtered informative locus, that is of maternal or fetal origin. The estimate is made by multiplying the ratio calculated in step 2007 by a scalar. The scalar represents a total expression level at the locus relative to expression of the second allele in the heterozygous individual.

**[0106]** To illustrate, if the second ("B") allele counted in the method is fetal-specific, then the B-allele ratio, calculated as above, represents the number of sequence reads for this allele as a fraction of all sequence reads for the filtered informative locus. Of the sequence reads containing the shared ("A") allele, some are contributed by the mother and some are contributed by the fetus. If the relative expression levels of the fetal-specific and shared alleles in the fetus are known, or can be estimated, then the B-allele ratio can be scaled to obtain an estimate of the fractional fetal contribution to all sequence reads for the locus. If the "A" and "B" alleles are expressed equally in the fetus, then the scalar is about two. The fractional maternal contribution to sequence reads at the locus is then one minus the fractional fetal contribution. Similar calculations can be made when the "B" allele at the filtered informative locus is maternal-specific.

### *B. Example: High fetal and maternal contributions*

**[0107]** Using a B-allele cutoff of 0.4 as described above, 0.91% of the circulating transcripts were found to show high contribution by the fetus during early gestation (i.e. the first and second trimesters). This percentage increased to 2.52% during late gestation (i.e. the third trimester). On the other hand, 42.58% and 50.98% showed high contribution by the mother during early and late gestations, respectively (**FIGS. 16, 17, and 18A**).

## V. USE OF A GENES FOR DIAGNOSING PREGNANCY-ASSOCIATED DISORDERS

### *A. Pregnancy-associated genes*

**[0108]** A method of identifying pregnancy-associated genes is also provided. The method includes receiving a plurality of first sequence reads and plurality of second sequence reads. The first sequence reads result from sequencing RNA molecules obtained from a blood plasma sample of a pregnant woman. The second sequence reads result from sequencing RNA molecules obtained from a blood plasma sample of a non-pregnant woman. The first sequence reads and the second sequence reads are aligned with a reference sequence, and a set of candidate genes is designated.

**[0109]** According to the method, sequence reads are then used to determine expression levels for each candidate gene in the samples from the pregnant woman and the non-pregnant woman. Specifically, for each candidate gene, a first number of transcripts corresponding to the candidate gene is determined using the first sequence reads; and a second number of transcripts corresponding to the candidate gene is determined using the second sequence reads. The first number of transcripts and the second number of transcripts can be normalized. A transcript ratio for the candidate gene is then calculated, where the transcript ratio includes the first number of transcripts divided by the second number of transcripts. The transcript ratio is then compared with a cutoff. The candidate gene is identified as a pregnancy-associated gene if the transcript ratio exceeds the cutoff.

**[0110]** In some embodiments of the method, normalizing the first number of transcripts corresponds to scaling the first number of transcripts by the total number of first sequence reads. Similarly, normalizing the second number of transcripts can correspond to scaling the second number of transcripts by the total number of second sequence reads. In other embodiments, normalizing the first number of transcripts for each candidate gene corresponds to scaling the first number of transcripts for the candidate gene by the total

number of first transcripts for all candidate genes. Normalizing the second number of transcripts for each candidate gene can correspond to scaling the second number of transcripts for the candidate gene by the total number of second transcripts for all candidate genes.

[0111] The method generally identifies genes as pregnancy-associated if they are expressed at higher levels in a pregnant woman as compared with a non-pregnant woman, all else being equal. The pregnant and non-pregnant women can be the same individual; that is, samples obtained from the individual before and after giving birth can be the sources of the first sequence reads and the second sequence reads, respectively.

[0112] We have shown that a subset of circulating RNA transcripts bearing fetal-specific alleles completely disappeared from the maternal plasma after delivery (**FIG. 18B**). These transcripts were considered to be fetal-specific in maternal plasma. On the other hand, a portion of maternal-specific alleles was also undetectable after delivery. Thus, we explored genes that showed up-regulation during pregnancy, termed pregnancy-associated genes, by directly comparing their representation in the pre- and post-delivery maternal plasma. We defined pregnancy-associated genes as those that were detected in the pre-delivery plasma of the third-trimester pregnant women, which their post-partum plasma level was decreased by $\geq$ 2-fold in both cases. By using a bioinformatics algorithm for data normalization and differential gene expression analysis, we compiled a list of 131 pregnancy-associated genes (**FIG. 21**). Among these genes, 15 were previously reported to be pregnancy-specific in maternal plasma [1,2,4,5,9,10,12]. Using one-step real-time RT-PCR, we have further validated the pregnancy-association of five newly-identified transcripts, which were abundant in pre-delivery maternal plasma, i.e. *STAT1*, *GBP1* and *HSD17B1* in 10 additional plasma samples from third-trimester pregnant women, as well as *KRT18* and *GADD45G* in 10 plasma samples from another cohort of third-trimester pregnant women (**FIG. 22**).

[0113] To assess the association of these 131 genes with pregnancy, hierarchical clustering was performed for all the plasma samples. A clear separation was observed between the plasma samples from pregnant women (i.e. early and late pregnancy) and those not associated with on-going pregnancy (i.e. non-pregnant controls and post-delivery) (**FIG. 23**).

[0114] Interestingly, when the expression patterns of these 131 genes were compared between the plasma samples of the two late pregnancy cases and their corresponding placenta and maternal blood cells, a closer resemblance was observed between the placenta and the

pre-delivery plasma samples, and between the maternal blood cells and the post-partum plasma samples (**FIG. 24A**). This observation supports the thesis that most pregnancy-associated genes are preferentially expressed in the placenta than in the maternal blood cells. Furthermore, the expression levels of these pregnancy-associated transcripts in the placentas and maternal plasma were positively correlated ($P < 0.05$, Spearman correlation) (**FIG. 24B**).

### *B. Differential gene expression in placenta and maternal blood*

**[0115]** While we were able to catalog a panel of pregnancy-associated genes through direct examination of the pre- and post-delivery maternal plasma samples, we have also mined the placental and blood cell RNA-seq data for comparison purposes. Assuming that pregnancy-associated genes should be those expressed at a high level in the placenta and at a low level in the maternal blood cells as reported before [10,15], we arbitrarily set a 20-fold difference as a minimum cut-off for the tissue-based analysis. This tissue-based analysis yielded a total of 798 potential candidate genes, in which the proportions of fetal and maternal contributions in the maternal plasma were calculated. A relatively high proportion of genes with predominant fetal contribution was identified (**FIG. 25**) when compared to that of the full transcriptome (**FIG. 18A**). However, the plasma-based strategy outperformed the tissue-based strategy in being able to identify a higher proportion of genes with predominant fetal contribution (**FIG. 25**).

### *C. Genes associated with disease or disorder*

**[0116]** Also provided herein is a method of identifying genes associated with a pregnancy-associated disorder. The method includes receiving a plurality of first sequence reads and plurality of second sequence reads. The first sequence reads result from sequencing RNA molecules obtained from a blood plasma sample of a healthy pregnant woman. The second sequence reads result from sequencing RNA molecules obtained from a blood plasma sample of a pregnant woman suffering from a pregnancy-associated disorder, or carrying a fetus suffering from a pregnancy-associated disorder. The first sequence reads and the second sequence reads are aligned with a reference sequence, and a set of candidate genes is designated.

**[0117]** According to the method, sequence reads are then used to determine expression levels for each candidate gene in the samples for the two pregnant women. Specifically, for each candidate gene, a first number of transcripts corresponding to the candidate gene is determined using the first sequence reads, and a second number of transcripts corresponding

to the candidate gene is determined using the second sequence reads. The first number of transcripts and the second number of transcripts can be normalized. A transcript ratio for the candidate gene is then calculated, where the transcript ratio includes the first number of transcripts divided by the second number of transcripts. The transcript ratio is then compared with a reference value. The candidate gene is identified as associated with the disorder if the transcript ratio deviates from the reference value.

**[0118]** In some embodiments of the method, normalizing the first number of transcripts corresponds to scaling the first number of transcripts by the total number of first sequence reads. Similarly, normalizing the second number of transcripts can correspond to scaling the second number of transcripts by the total number of second sequence reads. In other embodiments, normalizing the first number of transcripts for each candidate gene corresponds to scaling the first number of transcripts for the candidate gene by the total number of first transcripts for all candidate genes. Normalizing the second number of transcripts for each candidate gene can correspond to scaling the second number of transcripts for the candidate gene by the total number of second transcripts for all candidate genes.

**[0119]** According to the method of identifying genes associated with a pregnancy-associated disorder, in some embodiments, the reference value is 1. In some embodiments, the transcript ratio deviates from the reference value when the ratio of the transcript ratio and the reference value exceeds or falls below a cutoff. In some embodiments, the transcript ratio deviates from the reference value when the difference between the transcript ratio and the reference value exceeds a cutoff.

**[0120]** The method generally identifies a gene as associated with a pregnancy-associated disorder if the expression level of that gene differs significantly in pregnancies that do and do not demonstrate the disorder, all else being equal.

**[0121]** The diagnosis and monitoring of fetal disorders and pregnancy pathologies has previously been achieved by using maternal plasma RNAs that are associated with diseases. For example, the maternal plasma level of *corticotrophin releasing hormone (CRH)* mRNA has been shown to be useful for the noninvasive detection and prediction of preeclampsia [2,3]. The detection of *interleukin 1 receptor-like 1 (IL1RL1)* mRNA in maternal plasma has been demonstrated to be useful for identifying women with spontaneous preterm birth [8]. A panel of growth-related maternal plasma RNA markers had also been investigated for the noninvasive

assessment of fetal growth and intrauterine growth restriction [4]. In this study, we reason that novel disease-associated circulating RNA markers could be identified by directly compared the maternal plasma transcriptomes of normal pregnant women and women with complicated pregnancies, such as preeclampsia, intrauterine growth restriction, preterm labour and fetal aneuploidies. To demonstrate the feasibility of this approach, we performed RNA-Seq for maternal plasma samples obtained from three pregnant women who developed preeclampsia and seven uncomplicated pregnant women of matched gestations. We identified 98 transcripts that showed significant elevation in the plasma of preeclamptic pregnant women (**FIG. 26**). The newly identified preeclampsia-associated transcripts are potentially useful for prediction, prognostication and monitoring of the disease.

**[0122]** This technology can also be used for the prediction and monitoring of preterm labour. The technology can also be used for the prediction of imminent fetal demise. The technology can also be used for detecting diseases caused by gene mutations, as long as the gene concerned is transcribed in fetal or placental tissues and the transcripts are detectable in maternal plasma.

**[0123]** Plasma RNA-Seq can be applied in other clinical scenarios. For example, the plasma RNA-Seq methodology developed in this study would be potentially useful for studying other pathological conditions, such as cancer, in which aberrant plasma RNA concentrations have been reported [13,14]. For example, by comparing the plasma transcriptomes of a cancer patient before and after therapy, tumor-associated circulating RNA markers may be identified for noninvasive diagnostic application.

## VI. ALLELIC EXPRESSION PATTERNS OF SPECIFIC GENES

**[0124]** RNA-sequencing has been employed to examine the allelic expression pattern [25]. We postulated that the allelic expression pattern for a given gene would be retained when the RNA transcripts were released from the tissues into the circulation, hence could be detected in the plasma. In this study, we have examined the allelic counts of two RNA transcripts, namely *PAPPA*, a pregnancy-specific gene, and *H19*, an imprinted maternally expressed gene [26,27].

**[0125]** For the *PAPPA* gene, we analyzed a SNP, rs386088, which contains a fetal-specific SNP allele (**FIG. 27**). The absence of *PAPPA* RNA-seq reads in the post-delivery plasma samples indicated that it was indeed pregnancy-specific [4]. Of note, there was no statistically

significant difference in the proportions of fetal-allele read counts between the pre-delivery maternal plasma and the placental RNA samples ($P = 0.320$, $\chi^2$ test), indicating that the maternal plasma data reflected the bi-allelic expression pattern of *PAPPA* in the placenta.

**[0126]** We have recently reported that the DNA methylation pattern of the imprinted maternally expressed *H19* gene in the placenta and the maternal blood cells could be detected by bisulfite DNA sequencing of maternal plasma DNA [28]. Here, we have further examined whether the genomic imprinting status of *H19* gene could be explored at the RNA level. We first focused on a SNP site in exon 1 of the *H19* gene, rs2839698, which contains a maternal-specific allele, i.e. AA in the fetus and AG in the mother. As shown in **FIG. 28**, only G-allele was detected in the post-delivery maternal plasma (**FIG. 28**). Such monoallelic pattern was in accordance with its linkage to the unmethylated G-allele on the rs4930098 SNP site in the imprinting control region (**FIG. 29**). In the pre-delivery maternal plasma, while G-allele was present, A-allele, which was contributed by the placenta, was also detected (**FIG. 28**). In three other SNP sites, i.e. rs2839701, rs2839702 and rs3741219, which bear maternal-specific alleles, a similar allelic pattern was found, i.e. bi-allelic in the pre-delivery maternal plasma and monoallelic in the post-delivery maternal plasma (**FIG. 28**). It might be that the maternal-specific alleles on these SNP sites were in the same maternal haplotype, which was unmethylated and therefore being transcribed. Notably, *H19* RNA was not expressed in the maternal blood cells (**FIG. 28**), suggesting that the *H19* RNA molecules in the plasma were derived from maternal tissues/organs other than the blood cells. Non-placental and non-fetal tissues that have been reported to show *H19* expression included the adrenal gland, skeletal muscles, uterus, adipocytes, liver and pancreas [29].

## VII. DISCUSSION

**[0127]** In this work, we aimed to develop a technology for providing a global view of the transcriptomic activities in maternal plasma using RNA-seq. We have previously shown that the fractional fetal DNA in maternal plasma can be calculated by targeting one or several fetal-specific loci because the whole fetal genome is evenly represented in the maternal plasma [30]. Unlike circulating DNA, measurement of the proportion of fetal-derived RNA transcripts in the maternal plasma is less straightforward as it is complicated by differential gene expression in the fetal and maternal tissues and perhaps their release into the circulation. By performing RNA-seq on maternal plasma and examining the polymorphic differences

between the fetus and the mother, we were able to estimate the proportion of plasma transcripts contributed by the fetus. While maternal-derived transcripts dominated the plasma transcriptome, as one would anticipate, 3.70% and 11.28% of the circulating transcripts in the maternal plasma were derived from the fetus during early and late pregnancy, respectively. These fetal-derived transcripts include the RNA molecules co-contributed by the fetus and the mother, as well as those contributed solely by the fetus. We found the latter to constitute 0.90% and 2.52% of the maternal circulating transcripts, during early and late pregnancy, respectively. The higher representation of such fetal-specific genes during late pregnancy is perhaps correlated with an increase in the size of the fetus and the placenta as pregnancy progresses.

**[0128]** In this study, we have demonstrated that a balanced RNA allelic expression of the pregnancy-specific *PAPPA* gene in the placenta and the monoallelic expression of the imprinted maternally expressed *H19* gene could be observed in the maternal plasma. These data suggest that the maternal plasma could be used as a noninvasive sample source for the study of allele expression patterns.

**[0129]** By quantitative comparison of the RNA transcripts in the pre- and post-delivery maternal plasma samples, we have compiled a list of 131 genes that were up-regulated during pregnancy, as evident by their reduced representation in the post-partum plasma samples. As expected, the profile of these genes could be used to differentiate plasma samples of pregnant women from those of the non-pregnant women. Such direct comparison of the pre- and post-delivery maternal plasma samples has allowed us to, in a high-throughput manner, sort out pregnancy-associated genes, which may not necessarily be expressed at a much higher level in the placenta than in the maternal blood cells as demonstrated in previous work [10,15]. In essence, this direct plasma examination method presents another avenue for the discovery of circulating pregnancy-associated RNA transcripts, without an *a priori* knowledge of the transcriptomic profiles of the placental tissues and the blood cells.

**[0130]** While we have shown that RNA-seq is a feasible method to profile the plasma transcriptome, several technical issues can be further improved. First, the information yield for plasma RNA-seq could be increased by further optimization of the sequencing protocol, particularly in depletion of the highly transcribed rRNA and globin genes from the plasma. Second, we have focused only on the reference transcripts and have not yet explored individual isoforms. Future studies could include detection of novel transcripts and

differential analysis of the splicing variants and their isoforms [31-33] by increasing the sequencing read depth. Third, we have omitted ASE-filtering in the analysis of proportions of fetal- and maternal-derived transcripts for the early pregnancy samples as the chorionic villi and amniotic fluid had been exhausted for the genotyping analysis. Nonetheless, we have shown in late pregnancy that ASE-filtering had no pronounced impact on the identification of genes with predominant fetal and maternal contributions in the maternal plasma (**FIGS. 16 and 17**).

**[0131]** In conclusion, we have demonstrated that RNA-seq technology can be used to measure the proportion of fetal-derived transcripts and to identify circulating pregnancy-associated genes in the maternal plasma. This study has paved the path towards better comprehension of the transcriptomic landscape of maternal plasma hence facilitating the identification of biomarker candidates involved in pregnancy-related diseases. We envision that this technology would lead to new avenues for molecular diagnostics for pregnancy- or placenta-related diseases, and also for other diseases such as cancer [34].

## VIII. TRANSPLANTATION

**[0132]** As mentioned above, methods described herein can also be applied to transplantation. The methods for transplantation can be proceed in a similar manner as for the fetal analysis. For example, a genotype of the transplanted tissue can be obtained. Embodiments can identify loci where the transplanted tissue is heterozygous, and where the host organism (e.g., male or female) is homozygous. And, embodiments can identify loci where the transplanted tissue is homozygous, and where the host organism (e.g., male or female) is heterozygous. The same ratios can be computed, and compared to a cutoff value to determine whether a disorder exists.

## IX. COMPUTER SYSTEMS

**[0133]** Any of the computer systems mentioned herein may utilize any suitable number of subsystems. Examples of such subsystems are shown in **FIG. 30** in computer apparatus 3000. In some embodiments, a computer system includes a single computer apparatus, where the subsystems can be the components of the computer apparatus. In other embodiments, a computer system can include multiple computer apparatuses, each being a subsystem, with internal components.

[0134] The subsystems shown in **FIG. 30** are interconnected via a system bus 3075. Additional subsystems such as a printer 3074, keyboard 3078, storage device(s) 3079, monitor 3076, which is coupled to display adapter 3082, and others are shown. Peripherals and input/output (I/O) devices, which couple to I/O controller 3071, can be connected to the computer system by any number of means known in the art, such as serial port 3077. For example, serial port 3077 or external interface 3081 (e.g. Ethernet, Wi-Fi, etc.) can be used to connect computer system 3000 to a wide area network such as the Internet, a mouse input device, or a scanner. The interconnection via system bus 3075 allows the central processor 3073 to communicate with each subsystem and to control the execution of instructions from system memory 3072 or the storage device(s) 3079 (e.g., a fixed disk, such as a hard drive or optical disk), as well as the exchange of information between subsystems. The system memory 3072 and/or the storage device(s) 3079 may embody a computer readable medium. Any of the data mentioned herein can be output from one component to another component and can be output to the user.

[0135] A computer system can include a plurality of the same components or subsystems, e.g., connected together by external interface 3081 or by an internal interface. In some embodiments, computer systems, subsystem, or apparatuses can communicate over a network. In such instances, one computer can be considered a client and another computer a server, where each can be part of a same computer system. A client and a server can each include multiple systems, subsystems, or components.

[0136] It should be understood that any of the embodiments of the present invention can be implemented in the form of control logic using hardware (e.g. an application specific integrated circuit or field programmable gate array) and/or using computer software with a generally programmable processor in a modular or integrated manner. As user herein, a processor includes a multi-core processor on a same integrated chip, or multiple processing units on a single circuit board or networked. Based on the disclosure and teachings provided herein, a person of ordinary skill in the art will know and appreciate other ways and/or methods to implement embodiments of the present invention using hardware and a combination of hardware and software.

[0137] Any of the software components or functions described in this application may be implemented as software code to be executed by a processor using any suitable computer language such as, for example, Java, C++ or Perl using, for example, conventional or object-

oriented techniques. The software code may be stored as a series of instructions or commands on a computer readable medium for storage and/or transmission, suitable media include random access memory (RAM), a read only memory (ROM), a magnetic medium such as a hard-drive or a floppy disk, or an optical medium such as a compact disk (CD) or DVD (digital versatile disk), flash memory, and the like. The computer readable medium may be any combination of such storage or transmission devices.

**[0138]** Such programs may also be encoded and transmitted using carrier signals adapted for transmission via wired, optical, and/or wireless networks conforming to a variety of protocols, including the Internet. As such, a computer readable medium according to an embodiment of the present invention may be created using a data signal encoded with such programs. Computer readable media encoded with the program code may be packaged with a compatible device or provided separately from other devices (e.g., via Internet download). Any such computer readable medium may reside on or within a single computer product (e.g. a hard drive, a CD, or an entire computer system), and may be present on or within different computer products within a system or network. A computer system may include a monitor, printer, or other suitable display for providing any of the results mentioned herein to a user.

**[0139]** Any of the methods described herein may be totally or partially performed with a computer system including one or more processors, which can be configured to perform the steps. Thus, embodiments can be directed to computer systems configured to perform the steps of any of the methods described herein, potentially with different components performing a respective steps or a respective group of steps. Although presented as numbered steps, steps of methods herein can be performed at a same time or in a different order. Additionally, portions of these steps may be used with portions of other steps from other methods. Also, all or portions of a step may be optional. Additionally, any of the steps of any of the methods can be performed with modules, circuits, or other means for performing these steps.

**[0140]** The specific details of particular embodiments may be combined in any suitable manner without departing from the spirit and scope of embodiments of the invention. However, other embodiments of the invention may be directed to specific embodiments relating to each individual aspect, or specific combinations of these individual aspects.

**[0141]** The above description of exemplary embodiments of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive

or to limit the invention to the precise form described, and many modifications and variations are possible in light of the teaching above. The embodiments were chosen and described in order to best explain the principles of the invention and its practical applications to thereby enable others skilled in the art to best utilize the invention in various embodiments and with various modifications as are suited to the particular use contemplated.

**[0142]** A recitation of "a", "an" or "the" is intended to mean "one or more" unless specifically indicated to the contrary.

**[0143]** All patents, patent applications, publications, and descriptions mentioned here are incorporated by reference in their entirety for all purposes. None is admitted to be prior art.

## X. REFERENCES


1. Ng EKO, Tsui NBY, Lau TK, Leung TN, Chiu RWK, Panesar NS, Lit LC, Chan KW, Lo YMD. mRNA of placental origin is readily detectable in maternal plasma. Proc Natl Acad Sci U S A 2003;**100**: 4748-53.
2. Ng EKO, Leung TN, Tsui NBY, Lau TK, Panesar NS, Chiu RWK, Lo YMD. The concentration of circulating corticotropin-releasing hormone mRNA in maternal plasma is increased in preeclampsia. Clin Chem 2003;**49**: 727-31.
3. Farina A, Sekizawa A, Sugito Y, Iwasaki M, Jimbo M, Saito H, Okai T. Fetal DNA in maternal plasma as a screening variable for preeclampsia. A preliminary nonparametric analysis of detection rate in low-risk nonsymptomatic patients. Prenat Diagn 2004;**24**: 83-6.
4. Pang WW, Tsui MH, Sahota D, Leung TY, Lau TK, Lo YM, Chiu RW. A strategy for identifying circulating placental RNA markers for fetal growth assessment. Prenat Diagn 2009;**29**: 495-504.
5. Lo YMD, Tsui NBY, Chiu RWK, Lau TK, Leung TN, Heung MM, Gerovassili A, Jin Y, Nicolaides KH, Cantor CR, Ding C. Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection. Nat Med 2007;**13**: 218-23.
6. Tsui NBY, Akolekar R, Chiu RWK, Chow KCK, Leung TY, Lau TK, Nicolaides KH, Lo YMD. Synergy of total PLAC4 RNA concentration and measurement of the RNA single-nucleotide polymorphism allelic ratio for the noninvasive prenatal detection of trisomy 21. Clin Chem 2010;**56**: 73-81.
7. Tsui NBY, Wong BCK, Leung TY, Lau TK, Chiu RWK, Lo YMD. Non-invasive prenatal detection of fetal trisomy 18 by RNA-SNP allelic ratio analysis using maternal plasma SERPINB2 mRNA: a feasibility study. Prenat Diagn 2009;**29**: 1031-7.
8. Chim SS, Lee WS, Ting YH, Chan OK, Lee SW, Leung TY. Systematic identification of spontaneous preterm birth-associated RNA transcripts in maternal plasma. PLoS One 2012;7: e34328.
9. Wong BCK, Chiu RWK, Tsui NBY, Chan KCA, Chan LW, Lau TK, Leung TN, Lo YMD. Circulating placental RNA in maternal plasma is associated with a preponderance of 5' mRNA fragments: implications for noninvasive prenatal diagnosis and monitoring. Clin Chem 2005;**51**: 1786-95.

10. Tsui NBY, Chim SSC, Chiu RWK, Lau TK, Ng EKO, Leung TN, Tong YK, Chan KCA, Lo YMD. Systematic micro-array based identification of placental mRNA in maternal plasma: towards non-invasive prenatal gene expression profiling. J Med Genet 2004;41: 461-7.
11. Poon LL, Leung TN, Lau TK, Lo YMD. Presence of fetal RNA in maternal plasma. Clin Chem 2000;46:1832-4.
12. Go AT, Visser A, Mulders MA, Blankenstein MA, Van Vugt JM, Oudejans CB. Detection of placental transcription factor mRNA in maternal plasma. Clin Chem 2004;50:1413-4.
13. Smets EM, Visser A, Go AT, van Vugt JM, Oudejans CB. Novel biomarkers in preeclampsia. Clin Chim Acta 2006;364:22-32.
14. Purwosunu Y, Sekizawa A, Koide K, Farina A, Wibowo N, Wiknjosastro GH, et al. Cell-free mRNA concentrations of plasminogen activator inhibitor-1 and tissue-type plasminogen activator are increased in the plasma of pregnant women with preeclampsia. Clin Chem 2007;53:399-404.
15. Miura K, Miura S, Yamasaki K, Shimada T, Kinoshita A, Niikawa N, et al. The possibility of microarray-based analysis using cell-free placental mRNA in maternal plasma. Prenat Diagn 2010;30:849-61.
16. Ng EKO, El-Sheikhah A, Chiu RWK, Chan KC, Hogg M, Bindra R, et al. Evaluation of human chorionic gonadotropin beta-subunit mRNA concentrations in maternal serum in aneuploid pregnancies: a feasibility study. Clin Chem 2004;50:1055-7.
17. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621-8.
18. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 2008;321:956-60.
19. Kim J, Zhao K, Jiang P, Lu ZX, Wang J, Murray JC, Xing Y. Transcriptome landscape of the human placenta. BMC Genomics 2012;13:115.
20. Wang K, Li H, Yuan Y, Etheridge A, Zhou Y, Huang D, et al. The complex exogenous RNA spectra in human plasma: an interface with human gut biota? PLoS One 2012;7:e51009.
21. Li H, Guo L, Wu Q, Lu J, Ge Q, Lu Z. A comprehensive survey of maternal plasma miRNAs expression profiles using high-throughput sequencing. Clin Chim Acta 2012;413:568-76.
22. Williams Z, Ben-Dov IZ, Elias R, Mihailovic A, Brown M, Rosenwaks Z, Tuschl T. Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. Proc Natl Acad Sci U S A. 2013;110:4255-60.
23. Chim SSC, Shing TK, Hung EC, Leung TY, Lau TK, Chiu RWK, Lo YMD. Detection and characterization of placental microRNAs in maternal plasma. Clin Chem 2008;54:482-90.
24. Pickrell JK et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 2010;464:768-772.
25. Smith RM, Webb A, Papp AC, Newman LC, Handelman SK, Suhy A, et al. Whole transcriptome RNA-Seq allelic expression in human brain. BMC Genomics 2013;14:571.
26. Frost JM, Monk D, Stojilkovic-Mikic T, Woodfine K, Chitty LS, Murrell A, et al. Evaluation of allelic expression of imprinted genes in adult human blood. PLoS One 2010;5:e13556.

27. Daelemans C, Ritchie ME, Smits G, Abu-Amero S, Sudbery IM, Forrest MS, et al. High-throughput analysis of candidate imprinted genes and allele-specific gene expression in the human term placenta. BMC Genet 2010;11:25.
28. Lun FMF, Chiu RWK, Sun K, Leung TY, Jiang P, Chan KCA, et al. Noninvasive prenatal methylomic analysis by genomewide bisulfite sequencing of maternal plasma DNA. Clin Chem 2013;59:1583-94.
29. *Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, et al. BioGPS: an* extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 2009;10:R130.
30. Lo YMD, Chan KCA, Sun H, Chen EZ, Jiang P, Lun FMF, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med 2010;2:61ra91.
31. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011;25:1915-27.
32. St Laurent G, Shtokalo D, Tackett MR, Yang Z, Eremina T, Wahlestedt C, et al. Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. BMC Genomics 2012;13:504.
33. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome Res 2012;22:2008-17.
34. Fleischhacker M, Schmidt B. Circulating nucleic acids (CNAs) and cancer--a survey. Biochim Biophys Acta 2007;1775:181-232.

WHAT IS CLAIMED IS:

1. An ex vivo method of diagnosing a pregnancy-associated disorder using a sample from a female subject pregnant with a fetus, the method comprising:

receiving a plurality of reads, wherein the reads are obtained from an analysis of RNA molecules obtained from the sample, the sample containing a mixture of maternal- and fetal-derived RNA molecules;

identifying, by a computer system, locations of the reads in a reference sequence;

identifying one or more informative loci, each of which is homozygous in a first entity for a corresponding first allele and which is heterozygous in a second entity for the corresponding first allele and a corresponding second allele, wherein the first entity is the pregnant female subject or the fetus, and the second entity is the other one of the pregnant female subject and the fetus, wherein the one or more informative loci are identified using a sample from the female subject;

filtering the one or more informative loci to identify one or more filtered informative loci:

that are located within an expressed region of the reference sequence,

at which at least a first predetermined number of reads in the plurality of reads containing the corresponding first allele are located, and

at which at least a second predetermined number of reads in the plurality of reads containing the corresponding second allele are located,

for each of the filtered informative loci:

determining a first number of reads located at the filtered informative locus and containing the corresponding first allele,

determining a second number of reads located at the filtered informative locus and containing the corresponding second allele,

calculating a first sum of the first numbers and the second sum of the second numbers;

calculating a ratio of the first sum and the second sum, and

comparing the ratio to a cutoff value to determine whether the fetus, mother, or pregnancy has a pregnancy-associated disorder, the cutoff value being determined from one or more samples obtained from pregnant female subjects without the pregnancy-associated disorder.

2. The method of claim 1, wherein the sample of the pregnant woman is a sample of blood plasma.

3. The method of claim 1, wherein the second entity is the pregnant female subject.

4. The method of claim 1, wherein the second entity is the fetus.

5. The method of claim 1, wherein calculating a ratio of the first sum and the second sum comprises dividing the second sum by a sum of the first sum and the second sum.

6. The method of claim 1, wherein the analysis of RNA molecules obtained from the sample comprises sequencing the RNA molecules or cDNA copies thereof.

7. The method of claim 1, wherein the analysis of RNA molecules obtained from the sample comprises performing digital PCR.

8. The method of claim 1, wherein identifying locations of the reads in a reference sequence comprises aligning the reads to the reference sequence.

9. The method of claim 1, wherein the first predetermined number of reads is 1, and wherein the second predetermined number of reads is 1.

10. The method of claim 1, further comprising sequencing genomic DNA obtained from maternal tissue to determine a genotype of the pregnant female subject at each informative locus.

11. The method of claim 1, further comprising sequencing genomic DNA obtained from the placenta, chorionic villi, amniotic fluid, or maternal plasma to determine a genotype of the fetus at each informative locus.

12. The method of claim 1, further comprising estimating a portion of RNA in the sample that is of maternal or fetal origin, wherein the portion is the ratio multiplied by a scalar.

13. A method of determining, in a sample from a female subject pregnant with a fetus, a portion of RNA that is of fetal origin, the method comprising:

receiving a plurality of reads, wherein the reads are obtained from an analysis of RNA molecules obtained from the sample, the sample containing a mixture of maternal- and fetal-derived RNA molecules;

identifying, by a computer system, locations of the reads in a reference sequence;

identifying one or more informative maternal loci, each of which is homozygous in the fetus for a corresponding first allele and which is heterozygous in the pregnant female subject for the corresponding first allele and a corresponding second allele;

filtering the one or more informative maternal loci to identify one or more filtered informative maternal loci:

that are located within an expressed region of the reference sequence,

at which at least a first predetermined number of reads in the plurality of reads containing the corresponding first allele are located, and

at which at least a second predetermined number of reads in the plurality of reads containing the corresponding second allele are located,

for each of the filtered informative maternal loci:

determining a first number of reads located at the filtered informative maternal locus and containing the corresponding first allele,

determining a second number of reads located at the filtered informative maternal locus and containing the corresponding second allele,

calculating a sum of the first number and the second number,

calculating a maternal ratio of the second number divided by the sum,

determining a scalar representing a total expression at the filtered informative maternal locus relative to expression of the corresponding second allele in the pregnant female subject,

multiplying the maternal ratio by the scalar to obtain a maternal contribution,

calculating a fetal contribution as one minus the maternal contribution, and

determining a portion of RNA in the sample that is of fetal origin, wherein the portion is an average of the fetal contributions for the filtered informative maternal loci.

14. The method of claim 13, wherein the average is weighted by the sums for the filtered informative maternal loci.

15. The method of claim 13, wherein the scalar is assumed to be about 2 for a filtered informative maternal locus.

16. The method of claim 13, wherein the sample of the pregnant woman is a sample of blood plasma.

17. The method of claim 13, wherein the analysis of RNA molecules obtained from the sample comprises sequencing the RNA molecules or cDNA copies thereof.

18. The method of claim 13, wherein the analysis of RNA molecules obtained from the sample comprises performing digital PCR.

19. The method of claim 13, wherein identifying locations of the reads in a reference sequence comprises aligning the reads to the reference sequence.

20. The method of claim 13, wherein the first predetermined number of reads for the one or more filtered informative maternal loci is 1, and wherein the second predetermined number of reads for the one or more filtered informative maternal loci is 1.

21. The method of claim 13, further comprising sequencing genomic DNA obtained from maternal tissue to determine a genotype of the pregnant female subject at each informative maternal locus.

22. The method of claim 13, further comprising sequencing genomic DNA obtained from the placenta, chorionic villi, amniotic fluid, or maternal plasma to determine a genotype of the fetus at each informative maternal locus.

23. The method of claim 13, further comprising:

identifying one or more informative fetal loci, each of which is homozygous in the pregnant female subject for a corresponding first allele and which is heterozygous in the fetus for the corresponding first allele and a corresponding second allele;

filtering the one or more informative fetal loci to identify one or more filtered informative fetal loci:

that are located within an expressed region of the reference sequence,

at which at least a first predetermined number of reads in the plurality of reads containing the corresponding first allele are located, and

at which at least a second predetermined number of reads in the plurality of reads containing the corresponding second allele are located,

for each of the filtered informative fetal loci:

determining a first number of reads located at the filtered informative fetal locus and containing the corresponding first allele,

determining a second number of reads located at the filtered informative fetal locus and containing the corresponding second allele,

calculating a sum of the first number and the second number,

calculating a fetal ratio of the second number divided by the sum,

determining a scalar representing a total expression at the filtered informative fetal locus relative to expression of the corresponding second allele in the fetus;

multiplying the fetal ratio by the scalar to obtain a fetal contribution, and

determining a portion of RNA in the sample that is of fetal origin, wherein the portion is the average of the fetal contributions for the filtered informative maternal loci and the filtered informative fetal loci.

24. The method of claim 23, wherein the average is weighted by the sums for the filtered informative maternal loci and the filtered informative fetal loci.

25. The method of claim 23, wherein the scalar is assumed to be about 2 for a filtered informative fetal locus.

26. The method of claim 23, wherein the first predetermined number of reads for the one or more filtered informative fetal loci is 1 and the second predetermined number of reads for the one or more filtered informative fetal loci is 1.

27. The method of claim 23, further comprising sequencing genomic DNA obtained from maternal tissue to determine a genotype of the pregnant female subject at each informative fetal locus.

28. The method of claim 23, further comprising sequencing genomic DNA obtained from the placenta, chorionic villi, amniotic fluid, or maternal plasma to determine a genotype of the fetus at each informative fetal locus.

29. The method of claim 13 or 23, further comprising diagnosing a pregnancy-associated disorder by comparing the portion of RNA in the sample that is of fetal origin to a cutoff value.

30. A method of designating a genomic locus as a maternal or fetal marker by analyzing a sample from a female subject pregnant with a fetus, the method comprising:

receiving a plurality of reads, wherein the reads are obtained from an analysis of RNA molecules obtained from the sample, the sample containing a mixture of maternal- and fetal-derived RNA molecules;

identifying, by a computer system, locations of the reads in a reference sequence;

identifying one or more informative loci, each of which is homozygous in a first entity for a corresponding first allele and which is heterozygous in a second entity for the corresponding first allele and a corresponding second allele, wherein the first entity is the pregnant female subject or the fetus, and the second entity is the other one of the pregnant female subject and the fetus;

filtering the one or more informative loci to identify one or more filtered informative loci:

that are located within an expressed region of the reference sequence,

at which at least a first predetermined number of reads in the plurality of reads containing the corresponding first allele are located, and

at which at least a second predetermined number of reads in the plurality of reads containing the corresponding second allele are located,

for each of the filtered informative loci:

determining a first number of reads located at the filtered informative locus and containing the corresponding first allele,

determining a second number of reads located at the filtered informative locus and containing the corresponding second allele,

calculating a ratio of the first number and the second number,

designating the filtered informative locus as a marker for the second entity when the ratio exceeds a cutoff.

31. The method of claim 30, wherein the sample of the pregnant woman is a sample of blood plasma.

32. The method of claim 30, wherein the second entity is the pregnant female subject.

33. The method of claim 30, wherein the second entity is the fetus.

34. The method of claim 30, wherein calculating a ratio of the first number and the second number comprises dividing the second number the sum of the first number and the second number.

35. The method of claim 30, wherein the cutoff is from about 0.2 to about 0.5.

36. The method of claim 35, wherein the cutoff is 0.4.

37. The method of claim 30, wherein the analysis of RNA molecules obtained from the sample comprises sequencing the RNA molecules or cDNA copies thereof.

38. The method of claim 30, wherein the analysis of RNA molecules obtained from the sample comprises performing digital PCR.

39. The method of claim 30, wherein identifying locations of the reads in a reference sequence comprises aligning the reads to the reference sequence.

40. The method of claim 30, wherein the first predetermined number of reads is 1 and the second predetermined number of reads is 1.

41. The method of claim 30, further comprising sequencing genomic DNA obtained from maternal tissue to determine a genotype of the pregnant female subject at each informative locus.

42. The method of claim 30, further comprising sequencing genomic DNA obtained from the placenta, chorionic villi, amniotic fluid, or maternal plasma to determine a genotype of the fetus at each informative locus.

43. The method of claim 30, further comprising diagnosing a pregnancy-associated disorder based upon whether a filtered informative locus is designated as a marker for the second entity.

44. The method of claim 30, further comprising determining, for RNA in the sample that contains a filtered informative locus, a portion of the RNA that originates from the second entity, wherein the portion is determined by multiplying the ratio by a scalar.

45. The method of claim 44, wherein the scalar represents a total expression at the filtered informative maternal locus relative to expression of the corresponding second allele in the second entity.

46. The method of claim 44, wherein the scalar is assumed to be about 2.

47. The method of claim 44, further comprising determining a portion of the RNA that originates from the first entity by subtracting the portion of RNA that originates from the second entity from 1.

48. A system comprising one or more processors configured to perform the method of any one of claims 1 to 47.

49. A computer product comprising a computer readable medium storing a plurality of instructions for controlling a processor to perform the method of any one of claims 1 to 47.