NONINVASIVE DIAGNOSIS OF FETAL ANEUPLOIDY BY SEQUENCING

Inventors: Hei-Mun Christina Fan, Stephen R. Quake

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims priority from U.S. Provisional Patent Application No. 61/098,758, filed on September 20, 2008, which is hereby incorporated by reference in its entirety

STATEMENT OF GOVERNMENTAL SUPPORT

This invention was made with U.S. Government support under NIH Director's Pioneer Award DP1 OD000251. The U.S. Government has certain rights in this invention.

REFERENCE TO SEQUENCE LISTING, COMPUTER PROGRAM, OR COMPACT DISK

Applicants assert that the text copy of the Sequence Listing is identical to the Sequence Listing in computer readable form found on the accompanying computer file. Applicants incorporate the contents of the sequence listing by reference in its entirety.

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates to the field of molecular diagnostics, and more particularly to the field of prenatal genetic diagnosis.

Related Art

Presented below is background information on certain aspects of the present invention as they may relate to technical features referred to in the detailed description, but not necessarily described in detail. That is, certain components of the present invention may be described in greater detail in the materials discussed below. The discussion below should not be construed as an admission as to the relevance of the information to the claimed invention or the prior art effect of the material described.

Fetal aneuploidy and other chromosomal aberrations affect 9 out of 1000 live births (1). The gold standard for diagnosing chromosomal abnormalities is karyotyping of fetal cells obtained via invasive procedures such as chorionic villus sampling and amniocentesis. These

procedures impose small but potentially significant risks to both the fetus and the mother (2). Non-invasive screening of fetal aneuploidy using maternal serum markers and ultrasound are available but have limited reliability (3-5). There is therefore a desire to develop non-invasive genetic tests for fetal chromosomal abnormalities.

Since the discovery of intact fetal cells in maternal blood, there has been intense interest in trying to use them as a diagnostic window into fetal genetics (6-9). While this has not yet moved into practical application (10), the later discovery that significant amounts of cell-free fetal nucleic acids also exist in maternal circulation has led to the development of new non-invasive prenatal genetic tests for a variety of traits (11, 12). However, measuring aneuploidy remains challenging due to the high background of maternal DNA; fetal DNA often constitutes <10% of total DNA in maternal cell-free plasma (13).

Recently developed methods for aneuploidy rely on detection focus on allelic variation between the mother and the fetus. Lo et al. demonstrated that allelic ratios of placental specific mRNA in maternal plasma could be used to detect trisomy 21 in certain populations (14).

Similarly, they also showed the use of allelic ratios of imprinted genes in maternal plasma DNA to diagnose trisomy 18 (15). Dhallan et al. used fetal specific alleles in maternal plasma DNA to detect trisomy 21 (16). However, these methods are limited to specific populations because they depend on the presence of genetic polymorphisms at specific loci. We and others argued that it should be possible in principle to use digital PCR to create a universal, polymorphism independent test for fetal aneuploidy using maternal plasma DNA (17-19).

An alternative method to achieve digital quantification of DNA is direct shotgun sequencing followed by mapping to the chromosome of origin and enumeration of fragments per chromosome. Recent advances in DNA sequencing technology allow massively parallel sequencing (20), producing tens of millions of short sequence tags in a single run and enabling a deeper sampling than can be achieved by digital PCR. As is known in the art, the term "sequence tag" refers to a relatively short (e.g., 15-100) nucleic acid sequence that can be used to identify a certain larger sequence, e.g., be mapped to a chromosome or genomic region or gene. These can be ESTs or expressed sequence tags obtained from mRNA.

Specific Patents and Publications

Science 309:1476 (2 Sept. 2005) News Focus "An Earlier Look at Baby's Genes" describes attempts to develop tests for Down Syndrome using maternal blood. Early attempts to detect Down Syndrome using fetal cells from maternal blood were called "just modestly encouraging." The report also describes work by Dennis Lo to detect the Rh gene in a fetus where it is absent in the mother. Other mutations passed on from the father have reportedly been detected as well, such as cystic fibrosis, beta-thalassemia, a type of dwarfism and Huntington's disease. However, these results have not always been reproducible.

Venter et al., "The sequence of the human genome," *Science*, 2001 Feb 16;291(5507):1304-51 discloses the sequence of the human genome, which information is publicly available from NCBI. Another reference genomic sequence is a current NCBI build as obtained from the UCSC genome gateway.

Wheeler et al., "The complete genome of an individual by massively parallel DNA sequencing," *Nature*, 2008 Apr 17;452(7189):872-6 discloses the DNA sequence of a diploid genome of a single individual, James D. Watson, sequenced to 7.4-fold redundancy in two months using massively parallel sequencing in picolitre-size reaction vessels. Comparison of the sequence to the reference genome led to the identification of 3.3 million single nucleotide polymorphisms, of which 10,654 cause amino-acid substitution within the coding sequence.

Quake et al., US 2007/0202525 entitled "Non-invasive fetal genetic screening by digital analysis," published August 30, 2007, discloses a process in which maternal blood containing fetal DNA is diluted to a nominal value of approximately 0.5 genome equivalent of DNA per reaction sample.

Chiu et al., "Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic DNA sequencing of DNA in maternal plasma," *Proc. Natl. Acad. Sci.* 105(51):20458-20463 (December 23, 2008) discloses a method for determining fetal aneuploidy using massively parallel sequencing. Disease status determination (aneuploidy) was made by calculating a "z score." Z scores were compared with reference values, from a population restricted to euploid male fetuses. The authors noted in passing that G/C content affected the coefficient of variation.

Lo et al., "Diagnosing Fetal Chromosomal Aneuploidy Using Massively Parallel Genomic Sequencing," US 2009/0029377, published January 29, 2009, discloses a method in

which respective amounts of a clinically-relevant chromosome and of background chromosomes are determined from results of massively parallel sequencing. It was found that the percentage representation of sequences mapped to chromosome 21 is higher in a pregnant woman carrying a trisomy 21 fetus when compared with a pregnant woman carrying a normal fetus. For the four pregnant women each carrying a euploid fetus, a mean of 1.345% of their plasma DNA sequences were aligned to chromosome 21.

Lo et al., Determining a Nucleic Acid Sequence Imbalance," US 2009/0087847 published April 2, 2009, discloses a method for determining whether a nucleic acid sequence imbalance exists, such as an aneuploidy, the method comprising deriving a first cutoff value from an average concentration of a reference nucleic acid sequence in each of a plurality of reactions, wherein the reference nucleic acid sequence is either the clinically relevant nucleic acid sequence or the background nucleic acid sequence; comparing the parameter to the first cutoff value; and based on the comparison, determining a classification of whether a nucleic acid sequence imbalance exists.

BRIEF SUMMARY OF THE INVENTION

The following brief summary is not intended to include all features and aspects of the present invention, nor does it imply that the invention must include all features and aspects discussed in this summary.

The present invention comprises a method for analyzing a maternal sample, e.g., from peripheral blood. It is not invasive into the fetal space, as is amniocentesis or chorionic villi sampling. In the preferred method, fetal DNA which is present in the maternal plasma is used. The fetal DNA is in one aspect of the invention enriched due to the bias in the method towards shorter DNA fragments, which tend to be fetal DNA. The method is independent of any sequence difference between the maternal and fetal genome. The DNA obtained, preferably from a peripheral blood draw, is a mixture of fetal and maternal DNA. The DNA obtained is at least partially sequenced, in a method which gives a large number of short reads. These short reads act as sequence tags, in that a significant fraction of the reads are sufficiently unique to be mapped to specific chromosomes or chromosomal locations known to exist in the human genome. They are mapped exactly, or may be mapped with one mismatch, as in the examples below. By counting the number of sequence tags mapped to each chromosome (1-22, X and Y), the over- or under- representation of any chromosome or

chromosome portion in the mixed DNA contributed by an aneuploid fetus can be detected. This method does not require the sequence differentiation of fetal versus maternal DNA, because the summed contribution of both maternal and fetal sequences in a particular chromosome or chromosome portion will be different as between an intact, diploid chromosome and an aberrant chromosome, i.e., with an extra copy, missing portion or the like. In other words, the method does not rely on a priori sequence information that would distinguish fetal DNA from maternal DNA. The abnormal distribution of a fetal chromosome or portion of a chromosome (i.e., a gross deletion or insertion) may be determined in the present method by enumeration of sequence tags as mapped to different chromosomes. The median count of autosomal values (i.e., number of sequence tags per autosome) is used as a normalization constant to account for differences in total number of sequence tags is used for comparison between samples and between chromosomes The term "chromosome portion" is used herein to denote either an entire chromosome or a significant fragment of a chromosome. For example, moderate Down syndrome has been associated with partial trisomy 21q22.2→qter . By analyzing sequence tag density in predefined subsections of chromosomes (e.g., 10 to 100 kb windows), a normalization constant can be calculated, and chromosomal subsections quantified (e.g., 21q22.2). With large enough sequence tag counts, the present method can be applied to arbitrarily small fractions of fetal DNA. It has been demonstrated to be accurate down to 6% fetal DNA concentration. Exemplified below is the successful use of shotgun sequencing and mapping of DNA to detect fetal trisomy 21 (Down syndrome), trisomy 18 (Edward syndrome), and trisomy 13 (Patau syndrome), carried out non-invasively using cell-free fetal DNA in maternal plasma. This forms the basis of a universal, polymorphism-independent non-invasive diagnostic test for fetal aneuploidy. The sequence data also allowed us to characterize plasma DNA in unprecedented detail, suggesting that it is enriched for nucleosome bound fragments. The method may also be employed so that the sequence data obtained may be further analyzed to obtain information regarding polymorphisms and mutations.

Thus, the present invention comprises, in certain aspects, a method of testing for an abnormal distribution of a specified chromosome portion in a mixed sample of normally and abnormally distributed chromosome portions obtained from a single subject, such as a mixture of fetal and maternal DNA in a maternal plasma sample. One carries out sequence determinations on the DNA fragments in the sample, obtaining sequences from multiple chromosome portions of the mixed sample to obtain a number of sequence tags of sufficient

length of determined sequence to be assigned to a chromosome location within a genome and of sufficient number to reflect abnormal distribution. Using a reference sequence, one assigns the sequence tags to their corresponding chromosomes including at least the specified chromosome by comparing the sequence to reference genomic sequence. Often there will be on the order of millions of short sequence tags that are assigned to certain chromosomes, and, importantly, certain positions along the chromosomes. One then may determine a first number of sequence tags mapped to at least one normally distributed chromosome portion and a second number of sequence tags mapped to the specified chromosome portion, both chromosomes being in one mixed sample. The present method also involves correcting for nonuniform distribution sequence tags to different chromosomal portions. This is explained in detail below, where a number of windows of defined length are created along a chromosome, the windows being on the order of kilobases in length, whereby a number of sequence tags will fall into many of the windows and the windows covering each entire chromosome in question, with exceptions for non-informative regions, e.g., centromere regions and repetitive regions. Various average numbers, i.e., median values, are calculated for different windows and compared. By counting sequence tags within a series of predefined windows of equal lengths along different chromosomes, more robust and statistically significant results may be obtained. The present method also involves calculating a differential between the first number and the second number which is determinative of whether or not the abnormal distribution exists.

In certain aspects, the present invention may comprise a computer programmed to analyze sequence data obtained from a mixture of maternal and fetal chromosomal DNA. Each autosome (chr. 1-22) is computationally segmented into contiguous, non-overlapping windows. (A sliding window could also be used). Each window is of sufficient length to contain a significant number of reads (sequence tags, having about 20-100 bp of sequence) and not still have a number of windows per chromosome. Typically, a window will be between 10kb and 100kb, more typically between 40 and 60 kb. There would, then, for example, accordingly be approximately between 3,000 and 100,000 windows per chromosome. Windows may vary widely in the number of sequence tags that they contain, based on location (e.g., near a centromere or repeating region) or G/C content, as explained below. The median (i.e., middle value in the set) count per window for each chromosome is selected; then the median of the autosomal values is used to account for differences in total number of sequence tags obtained for different chromosomes and distinguish

interchromosomal variation from sequencing bias from aneuploidy. This mapping method may also be applied to discern partial deletions or insertions in a chromosome. The present method also provides a method for correcting for bias resulting from G/C content. For example, some the Solexa sequencing method was found to produce more sequence tags from fragments with increased G/C content. By assigning a weight to each sequence tag based on the G/C content of a window in which the read falls. The window for GC calculation is preferably smaller than the window for sequence tag density calculation.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a scatter plot graph showing sequence tag densities from eighteen samples, having five different genotypes, as indicated in the figure legend. Fetal aneuploidy is detectable by the over-representation of the affected chromosome in maternal blood. **Figure 1A** shows sequence tag density relative to the corresponding value of genomic DNA control; chromosomes are ordered by increasing G/C content. The samples shown as indicated, are plasma from a woman bearing a T21 fetus; plasma from a woman bearing a T18 fetus; plasma from a normal adult male; plasma from a woman bearing a normal fetus; plasma from a woman bearing a T13 fetus. Sequence tag densities vary more with increasing chromosomal G/C content. **Figure 1B** is a detail from **Fig. 1A**, showing chromosome 21 sequence tag density relative to the median chromosome 21 sequence tag density of the normal cases. Note that the values of 3 disomy 21 cases overlap at 1.0. The dashed line represents the upper boundary of the 99% confidence interval constructed from all disomy 21 samples. The chromosomes are listed in **Figure 1A** in order of G/C content, from low to high. This figure suggests that one would prefer to use as a reference chromosome in the mixed sample with a mid level of G/C content, as it can be seen that the data there are more tightly grouped. That is, chromosomes 18, 8, 2, 7, 12, 21 (except in suspected Down syndrome), 14, 9, and 11 may be used as the nominal diploid chromosome if looking for a trisomy. **Figure 1B** represents an enlargement of the chromosome 21 data.

Figure 2 is a scatter plot graph showing fetal DNA fraction and gestational age. The fraction of fetal DNA in maternal plasma correlates with gestational age. Fetal DNA fraction was estimated by three different ways: 1. From the additional amount of chromosomes 13, 18, and 21 sequences for T13, T18, and T21 cases respectively. 2. From the depletion in amount of chromosome X sequences for male cases. 3. From the amount of chromosome Y sequences present for male cases. The horizontal dashed line represents the estimated

minimum fetal DNA fraction required for the detection of aneuploidy. For each sample, the values of fetal DNA fraction calculated from the data of different chromosomes were averaged. There is a statistically significant correlation between the average fetal DNA fraction and gestational age (p=0.0051). The dashed line represents the simple linear regression line between the average fetal DNA fraction and gestational age. The R2 value represents the square of the correlation coefficient. **Figure 2** suggests that the present method may be employed at a very early stage of pregnancy. The data were obtained from the 10-week stage and later because that is the earliest stage at which chorionic villi sampling is done. (Amniocentesis is done later). From the level of the confidence interval, one would expect to obtain meaningful data as early as 4 weeks gestational age, or possibly earlier.

Figure 3 is a histogram showing size distribution of maternal and fetal DNA in maternal plasma. It shows the size distribution of total and chromosome Y specific fragments obtained from 454 sequencing of maternal plasma DNA from a normal male pregnancy. The distribution is normalized to sum to 1. The numbers of total reads and reads mapped to the Y-chromosome are 144992 and 178 respectively. Inset: Cumulative fetal DNA fraction as a function of sequenced fragment size. The error bars correspond to the standard error of the fraction estimated assuming the error of the counts of sequenced fragments follow Poisson statistics.

Figure 4 is a pair of line graphs showing distribution of sequence tags around transcription start sites (TSS) of ReSeq genes on all autosomes and chromosome X from plasma DNA sample of a normal male pregnancy (top, **Fig. 4A**) and randomly sheared genomic DNA control (bottom, **Fig. 4B**). The number of tags within each 5bp window was counted within ±1000bp region around each TSS, taking into account the strand each sequence tag mapped to. The counts from all transcription start sites for each 5bp window were summed and normalized to the median count among the 400 windows. A moving average was used to smooth the data. A peak in the sense strand represents the beginning of a nucleosome, while a peak in the anti-sense strand represents the end of a nucleosome. In the plasma DNA sample shown here, five well-positioned nucleosomes are observed downstream of transcription start sites and are represented as grey ovals. The number below within each oval represents the distance in base pairs between adjacent peaks in the sense and anti-sense strands, corresponding to the size of the inferred nucleosome. No obvious pattern is observed for the genomic DNA control.

Figure 5A is a scatter plot graph showing the mean sequence tag density for each chromosome of all samples, including cell-free plasma DNA from pregnant women and male donor, as well as genomic DNA control from male donor, is plotted above. Exceptions are chromosomes 13, 18 and 21, where cell-free DNA samples from women carrying aneuploid fetuses are excluded. The error bars represent standard deviation. The chromosomes are ordered by their G/C content. G/C content of each chromosome relative to the genome-wide value (41%) is also plotted. **Figure 5B** is a scatter plot of mean sequence tag density for each chromosome versus G/C content of the chromosome. The correlation coefficient is 0.927, and the correlation is statistically significant ($p<10^{-9}$).

Figure 5C is a scatter plot of the standard deviation of sequence tag density of each chromosome versus G/C content of the chromosome. The correlation coefficient between standard deviation of sequence tag density and the absolute deviation of chromosomal G/C content from the genome-wide G/C content is 0.963, and the correlation is statistically significant (p<10-12).

Figure 6 is a scatter plot graph showing percent difference of chromosome X sequence tag density of all samples as compared to the median chromosome X sequence tag density of all female pregnancies. All male pregnancies show under-representation of chromosome X.

Figure 7 is a scatter plot graph showing a comparison of the estimation of fetal DNA fraction for cell-free DNA samples from 12 male pregnancies using sequencing data from chromosomes X and Y. The dashed line represents a simple linear regression line, with a slope of 0.85. The R2 value represents the square of the correlation coefficient. There is a statistically significant correlation between fetal DNA fraction estimated from chromosomes X and Y (p=0.0015).

Figure 8 is a line graph showing length distribution of sequenced fragments from maternal cell-free plasma DNA sample of a normal male pregnancy at 1bp resolution. Sequencing was done on the 454/Roche platform. Reads that have at least 90% mapping to the human genome with greater than or equal to 90% accuracy are retained, totaling 144992 reads. Y-axis represents the number of reads obtained. The median length is 177bp while the mean length is 180bp.

Figure 9 is a schematic illustrating how sequence tag distribution is used to detect the over and under-representation of any chromosome, i.e., a trisomy (over representation) or a missing chromosome (typically an X or Y chromosome, since missing autosomes are generally lethal). As shown in left panels A and C, one first plots the number of reads obtained versus a window that is mapped to a chromosome coordinate that represents the position of the read along the chromosome. That is, chromosome 1 (panel A) can be seen to have about 2.8 x 108 bp. It would have this number divided by 50kb windows. These values are replotted (panels B and D) to show the distribution of the number of sequence tags/50kb window. The term "bin" is equivalent to a window. From this analysis, one can determine a median number of reads M for each chromosome, which, for purposes of illustration, may be observed along the x axis at the approximate center of the distribution and may be said to be higher if there are more sequence tags attributable to that chromosome. For chromosome 1, illustrated in panels A and B, one obtains a median M1. By taking the median M of all 22 autosomes, one obtains a normalization constant N that can be used to correct for differences in sequences obtained in different runs, as can be seen in Table 1. Thus, the normalized sequence tag density for chromosome 1 would be M1/N; for chromosome 22 it would be M22/N. Close examination of panel A, for example would show that towards the zero end of the chromosome, this procedure obtained about 175 reads per 50kb window. In the middle, near the centromere, there were no reads, because this portion of the chromosome is ill defined in the human genome library.

That is, in the left panels (A and C), one plots the distribution of reads per chromosome coordinate, i.e., chromosomal position in terms of number of reads within each 50kb non-overlapping sliding window. Then, one determines the distribution of the number of sequence tags for each 50 kb window, and obtains a median number of sequence tags per chromosome for all autosomes and chromosome X (Examples of chr 1 [top] and chr 22 [bottom] are illustrated here). These results are referred to as M. The median of the 22 values of M (from all autosomes, chromosomes 1 through 22) is used as the normalization constant N. The normalized sequence tag density of each chromosome is M/N (e.g., chr 1: M1/N; chr 22: M22/N). Such normalization is necessary to compare different patient samples since the total number of sequence tags (thus, the sequence tag density) for each patient sample is different (the total number of sequence tags fluctuates between ~8 to ~12 million). The analysis thus flows from frequency of reads per coordinate (A and C) to # reads per window (B and D) to a combination of all chromosomes.

Figure 10 is a scatter plot graph showing data from different samples, as in **Figure 1**, except that bias for G/C sampling has been eliminated.

Figure 11 is a scatter plot graph showing the weight given to different sequence samples according to percentage of G/C content, with lower weight given to samples with a higher G/C content. G/C content ranges from about 30% to about 70%; weight can range over a factor of about 3.

Figure 12 is a scatter plot graph which illustrates results of selected patients as indicated on the x axis, and, for each patient, a distribution of chromosome representation on the Y axis, as deviating from a representative t statistic, indicated as zero.

Figure 13 is a scatter plot graph showing the minimum fetal DNA percentage of which over- or under-representation of a chromosome could be detected with a 99.9% confidence level for chromosomes 21, 18, 13 and Chr. X, and a value for all other chromosomes.

Figure 14 is a scatter plot graph showing a linear relationship between log10 of minimum fetal DNA percentage that is needed versus log10 of the number of reads required.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Overview

Definitions

Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by those of ordinary skill in the art to which this invention belongs. Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present invention, the preferred methods and materials are described. Generally, nomenclatures utilized in connection with, and techniques of, cell and molecular biology and chemistry are those well known and commonly used in the art. Certain experimental techniques, not specifically defined, are generally performed according to conventional methods well known in the art and as described in various general and more specific references that are cited and discussed throughout the present specification. For purposes of the clarity, following terms are defined below.

"Sequence tag density" means the normalized value of sequence tags for a defined window of a sequence on a chromosome (in a preferred embodiment the window is about

50kb), where the sequence tag density is used for comparing different samples and for subsequent analysis. A "sequence tag" is a DNA sequence of sufficient length that it may be assigned specifically to one of chromosomes 1-22, X or Y. It does not necessarily need to be, but may be non-repetitive within a single chromosome. A certain, small degree of mismatch (0-1) may be allowed to account for minor polymorphisms that may exist between the reference genome and the individual genomes (maternal and fetal) being mapped. The value of the sequence tag density is normalized within a sample. This can be done by counting the number of tags falling within each window on a chromosome; obtaining a median value of the total sequence tag count for each chromosome; obtaining a median value of all of the autosomal values; and using this value as a normalization constant to account for the differences in total number of sequence tags obtained for different samples. A sequence tag density as calculated in this way would ideally be about 1 for a disomic chromosome. As further described below, sequence tag densities can vary according to sequencing artifacts, most notably G/C bias; this is corrected as described. This method does not require the use of an external standard, but, rather, provides an internal reference, derived from al of the sequence tags (genomic sequences), which may be, for example, a single chromosome or a calculated value from all autosomes.

"T21" means trisomy 21.

"T18" means trisomy 18.

"T13" means trisomy 13.

"Aneuploidy" is used in a general sense to mean the presence or absence of an entire chromosome, as well as the presence of partial chromosomal duplications or deletions or kilobase or greater size, as opposed to genetic mutations or polymorphisms where sequence differences exist.

"Massively parallel sequencing" means techniques for sequencing millions of fragments of nucleic acids, e.g., using attachment of randomly fragmented genomic DNA to a planar, optically transparent surface and solid phase amplification to create a high density sequencing flow cell with millions of clusters, each containing ~1,000 copies of template per sq. cm. These templates are sequenced using four-color DNA sequencing-by-synthesis technology. See, products offered by Illumina, Inc., San Diego, California. In the present work, sequences were obtained, as described below, with an Illumina/Solexa 1G Genome

Analyzer. The Solexa/Illumina method referred to below relies on the attachment of randomly fragmented genomic DNA to a planar, optically transparent surface. In the present case, the plasma DNA does not need to be sheared. Attached DNA fragments are extended and bridge amplified to create an ultra-high density sequencing flow cell with $\geq$ 50 million clusters, each containing ~1,000 copies of the same template. These templates are sequenced using a robust four-color DNA sequencing-by-synthesis technology that employs reversible terminators with removable fluorescent dyes. This novel approach ensures high accuracy and true base-by-base sequencing, eliminating sequence-context specific errors and enabling sequencing through homopolymers and repetitive sequences.

High-sensitivity fluorescence detection is achieved using laser excitation and total internal reflection optics. Short sequence reads are aligned against a reference genome and genetic differences are called using specially developed data analysis pipeline software.

Copies of the protocol for whole genome sequencing using Soelxa technology may be found at BioTechniques® Protocol Guide 2007 Published December 2006: p 29, www.biotechniques.com/default.asp?page=protocol&subsection=article_display&id=112378. Solexa's oligonucleotide adapters are ligated onto the fragments, yielding a fully-representative genomic library of DNA templates without cloning. Single molecule clonal amplification involves six steps: Template hybridization, template amplification, linearization, blocking 3' ends, denaturation and primer hybridization. Solexa's Sequencing-by-Synthesis utilizes four proprietary nucleotides possessing reversible fluorophore and termination properties. Each sequencing cycle occurs in the presence of all four nucleotides.

The presently used sequencing is preferably carried out without a preamplification or cloning step, but may be combined with amplification-based methods in a microfluidic chip having reaction chambers for both PCR and microscopic template-based sequencing. Only about 30 bp of random sequence information are needed to identify a sequence as belonging to a specific human chromosome. Longer sequences can uniquely identify more particular targets. In the present case, a large number of 25bp reads were obtained, and due to the large number of reads obtained, the 50% specificity enabled sufficient sequence tag representation.

Further description of a massively parallel sequencing method, which employed the below referenced 454 method is found in Rogers and Ventner, “Genomics: Massively parallel

sequencing," *Nature,* 437, 326-327 (15 September 2005). As described there, Rothberg and colleagues (Margulies, M. et al. *Nature* 437, 376–380 (2005)), have developed a highly parallel system capable of sequencing 25 million bases in a four-hour period — about 100 times faster than the current state-of-the-art Sanger sequencing and capillary-based electrophoresis platform. The method could potentially allow one individual to prepare and sequence an entire genome in a few days. The complexity of the system lies primarily in the sample preparation and in the microfabricated, massively parallel platform, which contains 1.6 million picoliter-sized reactors in a 6.4-cm^2 slide. Sample preparation starts with fragmentation of the genomic DNA, followed by the attachment of adaptor sequences to the ends of the DNA pieces. The adaptors allow the DNA fragments to bind to tiny beads (around 28 μ in diameter). This is done under conditions that allow only one piece of DNA to bind to each bead. The beads are encased in droplets of oil that contain all of the reactants needed to amplify the DNA using a standard tool called the polymerase chain reaction. The oil droplets form part of an emulsion so that each bead is kept apart from its neighbor, ensuring the amplification is uncontaminated. Each bead ends up with roughly 10 million copies of its initial DNA fragment. To perform the sequencing reaction, the DNA-template-carrying beads are loaded into the picoliter reactor wells — each well having space for just one bead. The technique uses a sequencing-by-synthesis method developed by Uhlen and colleagues, in which DNA complementary to each template strand is synthesized. The nucleotide bases used for sequencing release a chemical group as the base forms a bond with the growing DNA chain, and this group drives a light-emitting reaction in the presence of specific enzymes and luciferin. Sequential washes of each of the four possible nucleotides are run over the plate, and a detector senses which of the wells emit light with each wash to determine the sequence of the growing strand. This method has been adopted commercially by 454 Life Sciences.

Further examples of massively parallel sequencing are given in US 20070224613 by Strathmann, published September 27, 2007, entitled "Massively Multiplexed Sequencing." Also, for a further description of massively parallel sequencing, see US 2003/0022207 to Balasubramanian, et al., published January 30, 2003, entitled "Arrayed polynucleotides and their use in genome analysis."

General description of method and materials

Overview

Non-invasive prenatal diagnosis of aneuploidy has been a challenging problem because fetal DNA constitutes a small percentage of total DNA in maternal blood (13) and intact fetal cells are even rarer (6, 7, 9, 31, 32). We showed in this study the successful development of a truly universal, polymorphism-independent non-invasive test for fetal aneuploidy. By directly sequencing maternal plasma DNA, we could detect fetal trisomy 21 as early as 14th week of gestation. Using cell-free DNA instead of intact cells allows one to avoid complexities associated with microchimerism and foreign cells that might have colonized the mother; these cells occur at such low numbers that their contribution to the cell-free DNA is negligible (33, 34). Furthermore, there is evidence that cell-free fetal DNA clears from the blood to undetectable levels within a few hours of delivery and therefore is not carried forward from one pregnancy to the next (35-37).

Rare forms of aneuploidy caused by unbalanced translocations and partial duplication of a chromosome are in principle detectable by the approach of shotgun sequencing, since the density of sequence tags in the triplicated region of the chromosome would be higher than the rest of the chromosome. Detecting incomplete aneuploidy caused by mosaicism is also possible in principle but may be more challenging, since it depends not only on the concentration of fetal DNA in maternal plasma but also the degree of fetal mosaicism. Further studies are required to determine the effectiveness of shotgun sequencing in detecting these rare forms of aneuploidy.

The present method is applicable to large chromosomal deletions, such as 5p- Syndrome (five p minus), also known as Cat Cry Syndrome or Cri du Chat Syndrome. 5p- Syndrome is characterized at birth by a high-pitched cry, low birth weight, poor muscle tone, microcephaly, and potential medical complications. Similarly amenable disorders addressed by the present methods are p-, monosomy 9P, otherwise known as Alfi's Syndrome or 9P-, 22q11.2 deletion syndrome, Emanuel Syndrome, also known in the medical literature as the Supernumerary Der(22) Syndrome, trisomy 22, Unbalanced 11/22 Translocation or partial trisomy 11/22, Microdeletion and Microduplication at 16p11.2, which is associated with autism, and other deletions or imbalances, including those that are presently unknown.

An advantage of using direct sequencing to measure aneuploidy non-invasively is that it is able to make full use of the sample, while PCR based methods analyze only a few

targeted sequences. In this study, we obtained on average 5 million reads per sample in a single run, of which ~66,000 mapped to chromosome 21. Since those 5 million reads represent only a portion of one human genome, in principle less than one genomic equivalent of DNA is sufficient for the detection of aneuploidy using direct sequencing. In practice, a larger amount of DNA was used since there is sample loss during sequencing library preparation, but it may be possible to further reduce the amount of blood required for analysis.

Mapping shotgun sequence information (i.e., sequence information from a fragment whose physical genomic position is unknown) can be done in a number of ways, which involve alignment of the obtained sequence with a matching sequence in a reference genome. See, Li et al., "Mapping short DNA sequencing reads and calling variants using mapping quality score," *Genome Res.*, 2008 Aug 19. [Epub ahead of print].

We observed that certain chromosomes have large variations in the counts of sequenced fragments (from sample to sample, and that this depends strongly on the G/C content (**Figure 1A**) It is unclear at this point whether this stems from PCR artifacts during sequencing library preparation or cluster generation, the sequencing process itself, or whether it is a true biological effect relating to chromatin structure. We strongly suspect that it is an artifact since we also observe G/C bias on genomic DNA control, and such bias on the Solexa sequencing platform has recently been reported (38, 39). It has a practical consequence since the sensitivity to aneuploidy detection will vary from chromosome to chromosome; fortunately the most common human aneuploidies (such as 13, 18, and 21) have low variation and therefore high detection sensitivity. Both this problem and the sample volume limitations may possibly be resolved by the use of single molecule sequencing technologies, which do not require the use of PCR for library preparation (40).

Plasma DNA samples used in this study were obtained about 15 to 30 minutes after amniocentesis or chorionic villus sampling. Since these invasive procedures disrupt the interface between the placenta and maternal circulation, there have been discussions whether the amount of fetal DNA in maternal blood might increase following invasive procedures. Neither of the studies to date have observed a significant effect (41, 42).

Our results support this conclusion, since using the digital PCR assay we estimated that fetal DNA constituted less than or equal to 10% of total cell-free DNA in the majority of

our maternal plasma samples. This is within the range of previously reported values in maternal plasma samples obtained prior to invasive procedures (13). It would be valuable to have a direct measurement addressing this point in a future study.

The average fetal DNA fraction estimated from sequencing data is higher than the values estimated from digital PCR data by an average factor of two ($p<0.005$, paired t-test on all male pregnancies that have complete set of data). One possible explanation for this is that the PCR step during Solexa library preparation preferentially amplifies shorter fragments, which others have found to be enriched for fetal DNA (22, 23). Our own measurements of length distribution on one sample do not support this explanation, but nor can we reject it at this point. It should also be pointed out that using the sequence tags we find some variation of fetal fraction even in the same sample depending on which chromosome we use to make the calculation (**Figure 7**, Table 1). This is most likely due to artifacts and errors in the sequencing and mapping processes, which are substantial – recall that only half of the sequence tags map to the human genome with one error or less. Finally, it is also possible that the PCR measurements are biased since they are only sampling a tiny fraction of the fetal genome.

Our sequencing data suggest that the majority of cell-free plasma DNA is of apoptotic origin and shares features of nucleosomal DNA. Since nucleosome occupancy throughout the eukaryotic genome is not necessarily uniform and depends on factors such as function, expression, or sequence of the region (30, 43), the representation of sequences from different loci in cell-free maternal plasma may not be equal, as one usually expects in genomic DNA extracted from intact cells. Thus, the quantity of a particular locus may not be representative of the quantity of the entire chromosome and care must be taken when one designs assays for measuring gene dosage in cell-free maternal plasma DNA that target only a few loci.

Historically, due to risks associated with chorionic villus sampling and amniocentesis, invasive diagnosis of fetal aneuploidy was primarily offered to women who were considered at risk of carrying an aneuploid fetus based on evaluation of risk factors such as maternal age, levels of serum markers, and ultrasonographic findings. Recently, an American College of Obstetricians and Gynecologists (ACOG) Practice Bulletin recommended that "invasive diagnostic testing for aneuploidy should be available to all women, regardless of maternal age" and that "pretest counseling should include a discussion of the risks and benefits of invasive testing compared with screening tests"(2).

A noninvasive genetic test based on the results described here and in future large-scale studies would presumably carry the best of both worlds: minimal risk to the fetus while providing true genetic information. The costs of the assay are already fairly low; the sequencing cost per sample as of this writing is about $700 and the cost of sequencing is expected to continue to drop dramatically in the near future.

Shotgun sequencing can potentially reveal many more previously unknown features of cell-free nucleic acids such as plasma mRNA distributions, as well as epigenetic features of plasma DNA such as DNA methylation and histone modification, in fields including perinatology, oncology and transplantation, thereby improving our understanding of the basic biology of pregnancy, early human development and disease.

Sequencing Methods

Commercially available sequencing equipment was used in the present illustrative examples, namely the Solexa/Illumina sequencing platform and the 454/Roche platform. It will be apparent to those skilled in the art that a number of different sequencing methods and variations can be used. One sequencing method that can be used to advantage in the present methods involves paired end sequencing. Fluorescently labeled sequencing primers could be used to simultaneously sequence both strands of a dsDNA template, as described e.g., in Wiemann et al. (*Anal. Biochem.* 224: 117 [1995]; *Anal. Biochem.* 234: 166 [1996]. Recent examples of this technique have demonstrated multiplex co-sequencing using the four-color dye terminator reaction chemistry pioneered by Prober et al. (*Science* 238: 336 [1987]). Solexa/Illumina offers a "Paired End Module" to its Genome Analyzer. Using this module, after the Genome Analyzer has completed the first sequencing read, the Paired-End Module directs the resynthesis of the original templates and the second round of cluster generation. The Paired-End Module is connected to the Genome Analyzer through a single fluidic connection. In addition, 454 has developed a protocol to generate a library of Paired End reads. These Paired End reads are approximately 84-nucleotide DNA fragments that have a 44-mer adaptor sequence in the middle flanked by a 20-mer sequence on each side. The two flanking 20-mers are segments of DNA that were originally located approximately 2.5 kb apart in the genome of interest.

By using paired end reads in the present method, one may obtain more sequence information from a given plasma DNA fragment, and, significantly, one may also obtain sequence information from both ends of the fragment. The fragment is mapped to the human

genome as explained here elsewhere. After mapping both ends, one may deduce the length of the starting fragment. Since fetal DNA is known to be shorter than maternal DNA fragments circulating in plasma, one may use this information about the length of the DNA fragment to effectively increase the weight given to sequences obtained from shorter (e.g., about 300 bp or less) DNA fragments. Methods for weighting are given below.

Another method for increasing sensitivity to fetal DNA is to focus on certain regions within the human genome. One may use sequencing methods which select a priori sequences which map to the chromosomes of interest (as described here elsewhere, such as 18, 21, 13, X and Y). One may also choose to focus, using this method, on partial chromosomal deletions, such as 22q11 deletion syndrome. Other microdeletions and microduplications are set forth in Table 1 of US 2005/0181410, published Aug. 18 2005 under the title "Methods and apparatuses for achieving precision genetic diagnosis."

In sequencing selected subsequences, one may employ sequence-based methodologies such as sequencing by array, or capture beads with specific genomic sequences used as capture probes. The use of a sequencing array can be implemented as described in Chetverin et al., "Oligonucleotide arrays: new concepts and possibilities," *Biotechnology* (N Y). 1994 Nov;12(11):1093-9, as well as Rothberg, US 2002/0012930 A1 entitled "Method of Sequencing a Nucleic Acid," and Reeve et al., "Sequencing by Hybridization," US 6,399,364. In these methods, the target nucleic acid to be sequenced may be genomic DNA, cDNA or RNA. The sample is rendered single stranded and captured under hybridizing conditions with a number of single stranded probes which are catalogued by bar coding or by physical separation in an array. Emulsion PCR, as used in the 454 system, the SOLiD system, and Polonator (Dover Systems) and others may also be used, where capture is directed to specific target sequences, e.g., genome sequences mapping uniquely to chromosome 21 or other chromosome of interest, or to a chromosome region such as 15q11 (Prader-Willi syndrome), or excessive CGG repeats in the FMR1 gene (fragile X syndrome).

The subsequencing method is in one aspect contrary to conventional massively parallel sequencing methodologies, which seek to obtain all of the sequence information in a sample. This alternative method selectively ignores certain sequence information by using a sequencing method which selectively captures sample molecules containing certain predefined sequences. One may also use the sequencing steps exactly as exemplified, but in mapping the sequence fragments obtained, give greater weight to sequences which map to

areas known to be more reliable in their coverage, such as exons. Otherwise, the method proceeds as described below, where one obtains a large number of sequence reads from one or more reference chromosomes, which are compared to a large number of reads obtained from a chromosome of interest, after accounting for variations arising from chromosomal length, G/C content, repeat sequences and the like.

One may also focus on certain regions within the human genome according to the present methods in order to identify partial monosomies and partial trisomies. As described below, the present methods involve analyzing sequence data in a defined chromosomal sliding "window," such as contiguous, nonoverlapping 50Kb regions spread across a chromosome. Partial trisomies of 13q, 8p (8p23.1), 7q, distal 6p, 5p, 3q (3q25.1), 2q, 1q (1q42.1 and 1q21-qter), partial Xpand monosomy 4q35.1 have been reported, among others. For example, partial duplications of the long arm of chromosome 18 can result in Edwards syndrome in the case of a duplication of 18q21.1-qter (See, Mewar et al., "Clinical and molecular evaluation of four patients with partial duplications of the long arm of chromosome 18," *Am J Hum Genet.* 1993 Dec;53(6):1269-78).

Shotgun Sequencing of Cell-free Plasma DNA

Cell-free plasma DNA from 18 pregnant women and a male donor, as well as whole blood genomic DNA from the same male donor, were sequenced on the Solexa/Illumina platform. We obtained on average ~10 million 25bp sequence tags per sample. About 50% (i.e., ~5 million) of the reads mapped uniquely to the human genome with at most 1 mismatch against the human genome, covering ~4% of the entire genome. An average of ~154,000, ~135,000, ~66,000 sequence tags mapped to chromosomes 13, 18, and 21, respectively. The number of sequence tags for each sample is detailed in the following Table 1 and Table 2.

Table 1.

Sample	Fetal Karyotype	Gestational Age (weeks)	Volume of Plasma	Amt of DNA	Approx Amount of Input DNA *	Total Number of Sequence Tags
P1 Plasma DNA§	47XX +21	35	1.6	761	8.0	8206694
P2 Plasma DNA§	47XY +21	18	1.4	585	5.2	7751384
P6 Plasma DNA§	47XX +21	14	1.6	410	4.3	6699183
P7 Plasma DNA§	47XY +21	18	2.2	266	3.8	8324473
P14 Plasma DNA§	47XX +21	23	3.2	57	1.2	8924944
P17 Plasma DNA§	47XX +21	16	2.3	210	3.2	11599833
P19 Plasma DNA§	46XY	18	3.2	333	7.0	7305417
P20 Plasma DNA§	47XY +21	18	1.3	408	3.6	11454876
P23 Plasma DNA§	46XY	10	1.6	258	2.7	11851612
P26 Plasma DNA§	46XY	13	3.0	340	6.7	11471297
P31 Plasma DNA§	46XY	20	2.2	278	4.0	8967562
P40 Plasma DNA§	46XY	11	2.6	217	3.7	9205197
P42 Plasma DNA§	46XY	11	3.0	276	5.5	8364774
P52 Plasma DNA§	47XY +21	25	1.6	645	6.8	9192596
P53 Plasma DNA§	47XX +21	19	1.6	539	5.7	9771887
P57 Plasma DNA§	47XX +18	23	2.0	199	2.6	15041417
P59 Plasma DNA§	47XY +18	21	2.0	426	5.6	11910483
P64 Plasma DNA§	47XY +13	17	1.8	204	2.4	12097478
Male Donor Plasma DNA§	-	-	1.8	485	5.8	6669125
Male Donor Whole Blood Genomic DNA§	-	-	-	-	2.1	8519495
P25 Plasma DNA¶	46XY	11	5.6	132	4.9	242599
P13 Plasma DNA§	46XY	18	5.6	77	2.9	4168455

Table 2.

Sample	Number of Sequence Tags Mapped Uniquely to the Human Genome (hg18) with At Most 1 Mismatch	% Fetal DNA Estimated By Digital PCR with SRY Assay (male fetuses)	% Fetal DNA Estimated by ChrY Sequence Tags (male fetuses)	% Fetal DNA Estimated by Depletion of ChrX Sequence Tags (male fetuses)	% Fetal DNA Estimated by Addition of Trisomic Chromosome Sequence Tags (aneuploid fetuses)	Overall G/C content Of Sequence Tags (%)
P1 Plasma DNA[§]	4632637	-	-	-	35.0	43.65
P2 Plasma DNA[§]	4313884	6.4	15.4	21.6	15.5	48.72
P6 Plasma DNA[§]	3878383	-	-	-	22. 9	44.78
P7 Plasma DNA[§]	4294865	9.1	31.0	33.8	28.6	48.07
P14 Plasma DNA[§]	3603767	-	-	-	30.5	46.38
P17 Plasma DNA[§]	5968932	-	-	-	7.8	44.29
P19 Plasma DNA[§]	3280521	<5.9[‡]	4.14	21.5	-	50.09
P20 Plasma DNA[§]	6032684	10.0	15.7	11.3	11.5	44.02
P23 Plasma DNA[§]	6642795	5.3	12.2	9.6	-	43.80
P26 Plasma DNA[§]	3851477	10.3	18.2	14.2	-	42.51
P31 Plasma DNA[§]	4683777	Missing data[‡]	13.2	17.0	-	48.27
P40 Plasma DNA[§]	4187561	8.6	20.0	17.1	-	42.65
P42 Plasma DNA[§]	4315527	<4.4[‡]	9.7	7.9	-	44.14
P52 Plasma DNA[§]	5126837	6.3	25.0	26.3	26.4	44.34
P53 Plasma DNA[§]	5434222	-	-	-	25.8	44.18
P57 Plasma DNA[§]	7470487	-	-	-	23.0	42.89
P59 Plasma DNA[§]	6684871	26.4	44.0	39.8	45.1	43.64
P64 Plasma	6701148	<4.4[‡]	14.0	8.9	16.7	44.21

DNA[§]						
Male Donor Plasma DNA[§]	3692931	-	-	-	-	48.30
Male Donor Whole Blood Genomic DNA[§]	5085412	-	-	-	-	46.53
P25 Plasma DNA[¶]	144992[†]	-	-	-	-	41.38
P13 Plasma DNA[§]	2835333	9.8	5.7	n/a[‖]	-	39.60

The volume of plasma is the volume used for Sequencing Library Creation (ml). The amount of DNA is in Plasma (cell equivalent/ml plasma)*. The approximate amount of input DNA is that use for Sequencing Library Construction (ng).

---

*As quantified by digital PCR with EIF2C1 Taqman Assay, converting from copies to ng assuming 6.6pg/cell equivalent.

[†]For 454 sequencing, this number represents the number of reads with at least 90% accuracy and 90% coverage when mapped to hg18.

[‡]Insufficient materials were available for quantifying fetal DNA % with digital PCR for these samples (either no samples remained for analysis or there was insufficient sampling).

[§]Sequenced on Solexa/Illumina platform; [¶]Sequenced on 454/Roche platform

[‖]Sample P13 was the first to be analyzed by shotgun sequencing. It was a normal fetus and the chromosome value was clearly disomic. However, there were some irregularities with this sample and it was not included in further analysis. This sample was sequenced on a different Solexa instrument than the rest of the samples of this study, and it was sequenced in the presence of a number of samples of unknown origin. The G/C content of this sample was lower than the G/C bias of the human genome, while the rest of the samples are above. It had the lowest number of reads, and also the smallest number of reads mapped successfully to the human genome. This sample appeared to be outlier in sequence tag density for most chromosomes and the fetal DNA fraction calculated from chromosomes X was not well defined. For these reasons we suspect that the irregularities are due to technical problems with the sequencing process.

In Table 1 and Table 2, each sample represents a different patient, e.g., P1 in the first row. The total number of sequence tags varied but was frequently was in the 10 million range, using the Solexa technology. The 454 technology used for P25 and P13 gave a lower number of reads.

We observed a non-uniform distribution of sequence tags across each chromosome. This pattern of intra-chromosomal variation was common among all samples, including randomly sheared genomic DNA, indicating the observed variation was most probably due to sequencing artifacts. We applied an arbitrary sliding window of 50kb across each chromosome and counted the number of tags falling within each window. The window can be varied in size to account for larger numbers of reads (in which cases a smaller window, e.g., 10 kb, gives a more detailed picture of a chromosome) or a smaller number of reads, in which case a larger window (e.g., 100kb) may still be used and will detect gross chromosome deletions, omissions or duplications. The median count per 50kb window for each chromosome was selected. The median of the autosomal values (i.e., 22 chromosomes) was used as a normalization constant to account for the differences in total number of sequence tags obtained for different samples. The inter-chromosomal variation within each sample was also consistent among all samples (including genomic DNA control). The mean sequence tag density of each chromosome correlates with the G/C content of the chromosome ($p<10^{-9}$) (**Figure 5A, 5B**). The standard deviation of sequence tag density for each chromosome also correlates with the absolute degree of deviation in chromosomal G/C content from the genome-wide G/C content ($p<10^{-12}$) (**Figure 5A, 5C**). The G/C content of sequenced tags of all samples (including the genomic DNA control) was on average 10% higher than the value of the sequenced human genome (41%) (21)(Table 2), suggesting that there is a strong G/C bias stemming from the sequencing process. We plotted in **Figure 1A** the sequence tag density for each chromosome (ordered by increasing G/C content) relative to the corresponding value of the genomic DNA control to remove such bias.

Detection of Fetal Aneuploidy

The distribution of chromosome 21 sequence tag density for all 9 T21 pregnancies is clearly separated from that of pregnancies bearing disomy 21 fetuses ($p<10^{-5}$), Student's t-test) (**Figure 1A** and **1B**). The coverage of chromosome 21 for T21 cases is about ~4-18% higher (average ~11%) than that of the disomy 21 cases. Because the sequence tag density of chromosome 21 for T21 cases should be $(1+\varepsilon/2)$ of that of disomy 21 pregnancies, where ε is

the fraction of total plasma DNA originating from the fetus, such increase in chromosome 21 coverage in T21 cases corresponds to a fetal DNA fraction of ~8% – 35% (average ~23%) (Table 1, **Figure 2**). We constructed a 99% confidence interval of the distribution of chromosome 21 sequence tag density of disomy 21 pregnancies. The values for all 9 T21 cases lie outside the upper boundary of the confidence interval and those for all 9 disomy 21 cases lie below the boundary (**Figure 1B**). If we used the upper bound of the confidence interval as a threshold value for detecting T21, the minimum fraction of fetal DNA that would be detected is ~2%.

Plasma DNA of pregnant women carrying T18 fetuses (2 cases) and a T13 fetus (1 case) were also directly sequenced. Over-representation was observed for chromosome 18 and chromosome 13 in T18 and T13 cases respectively (**Figure 1A**). While there were not enough positive samples to measure a representative distribution, it is encouraging that all of these three positives are outliers from the distribution of disomy values. The T18 are large outliers and are clearly statistically significant ($p<10^{-7}$), while the statistical significance of the single T13 case is marginal ($p<0.05$). Fetal DNA fraction was also calculated from the over-represented chromosome as described above (**Figure 2**, Table 1).

Fetal DNA Fraction in Maternal Plasma

Using digital Taqman PCR for a single locus on chromosome 1, we estimated the average cell-free DNA concentration in the sequenced maternal plasma samples to be ~360 cell equivalent/ml of plasma (range: 57 to 761 cell equivalent/ml plasma) (Table 1), in rough accordance to previously reported values (13). The cohort included 12 male pregnancies (6 normal cases, 4 T21 cases, 1 T18 case and 1 T13 case) and 6 female pregnancies (5 T21 cases and 1 T18 case). DYS14, a multi-copy locus on chromosome Y, was detectable in maternal plasma by real-time PCR in all these pregnancies but not in any of the female pregnancies (data not shown). The fraction of fetal DNA in maternal cell-free plasma DNA is usually determined by comparing the amount of fetal specific locus (such as the SRY locus on chromosome Y in male pregnancies) to that of a locus on any autosome that is common to both the mother and the fetus using quantitative real-time PCR (13, 22, 23). We applied a similar duplex assay on a digital PCR platform (see Methods) to compare the counts of the SRY locus and a locus on chromosome 1 in male pregnancies. SRY locus was not detectable in any plasma DNA samples from female pregnancies. We found with digital PCR that for

the majority samples, fetal DNA constituted §10% of total DNA in maternal plasma (Table 2), agreeing with previously reported values (13).

The percentage of fetal DNA among total cell-free DNA in maternal plasma can also be calculated from the density of sequence tags of the sex chromosomes for male pregnancies. By comparing the sequence tag density of chromosome Y of plasma DNA from male pregnancies to that of adult male plasma DNA, we estimated fetal DNA percentage to be on average ~ 19% (range: 4-44%) for all male pregnancies (Table 2, above, **Figure 2**). Because human males have 1 fewer chromosome X than human females, the sequence tag density of chromosome X in male pregnancies should be (1-e/2) of that of female pregnancies, where e is fetal DNA fraction. We indeed observed under-representation of chromosome X in male pregnancies as compared to that of female pregnancies (**Figure 5**). Based on the data from chromosome X, we estimated fetal DNA percentage to be on average ~19% (range: 8-40%) for all male pregnancies (Table 2, above, **Figure 2**). The fetal DNA percentage estimated from chromosomes X and Y for each male pregnancy sample correlated with each other (p=0.0015) (**Figure 7**).

We plotted in **Figure 2** the fetal DNA fraction calculated from the over-representation of trisomic chromosome in aneuploid pregnancies, and the under-representation of chromosome X and the presence of chromosome Y for male pregnancies against gestational age. The average fetal DNA fraction for each sample correlates with gestational age (p=0.0051), a trend that is also previously reported (13).

Size Distribution of Cell-Free Plasma DNA

We analyzed the sequencing libraries with a commercial lab-on-a-chip capillary electrophoresis system. There is a striking consistency in the peak fragment size, as well as the distribution around the peak, for all plasma DNA samples, including those from pregnant women and male donor. The peak fragment size was on average 261bp (range: 256-264bp). Subtracting the total length of the Solexa adaptors (92bp) from 260bp gives 169bp as the actual peak fragment size. This size corresponds to the length of DNA wrapped in a chromatosome, which is a nucleosome bound to a H1 histone (24). Because the library preparation includes an 18-cycle PCR, there are concerns that the distribution might be biased. To verify that the size distribution observed in the electropherograms is not an artifact of PCR, we also sequenced cell-free plasma DNA from a pregnant woman carrying a male fetus using the 454 platform. The sample preparation for this system uses emulsion PCR,

which does not require competitive amplification of the sequencing libraries and creates product that is largely independent of the amplification efficiency. The size distribution of the reads mapped to unique locations of the human genome resembled those of the Solexa sequencing libraries, with a predominant peak at 176bp, after subtracting the length of 454 universal adaptors (**Figure 3** and **Figure 8**). These findings suggest that the majority of cell-free DNA in the plasma is derived from apoptotic cells, in accordance with previous findings (22, 23, 25, 26).

Of particular interest is the size distribution of maternal and fetal DNA in maternal cell-free plasma. Two groups have previously shown that the majority of fetal DNA has size range of that of mono-nucleosome (<200-300bp), while maternal DNA is longer. Because 454 sequencing has a targeted read-length of 250bp, we interpreted the small peak at around 250bp (**Figure 3** and **Figure 8**) as the instrumentation limit from sequencing higher molecular weight fragments. We plotted the distribution of all reads and those mapped to Y-chromosome (**Figure 3**). We observed a slight depletion of Y-chromosome reads in the higher end of the distribution. Reads <220bp constitute 94% of Y-chromosome and 87% of the total reads. Our results are not in complete agreement with previous findings in that we do not see as dramatic an enrichment of fetal DNA at short lengths (22, 23). Future studies will be needed to resolve this point and to eliminate any potential residual bias in the 454 sample preparation process, but it is worth noting that the ability to sequence single plasma samples permits one to measure the distribution in length enrichments across many individual patients rather than measuring the average length enrichment of pooled patient samples.

Cell-Free Plasma DNA Shares Features of Nucleosomal DNA

Since our observations of the size distribution of cell-free plasma DNA suggested that plasma DNA is mainly apoptotic of origin, we investigated whether features of nucleosomal DNA and positioning are found in plasma DNA. One such feature is nucleosome positioning around transcription start sites. Experimental data from yeast and human have suggested that nucleosomes are depleted in promoters upstream of transcription start sites and nucleosomes are well-positioned near transcription start sites (27-30). We applied a 5bp window spanning +/- 1000bp of transcription start sites of all RefSeq genes and counted the number of tags mapping to the sense and antisense strands within each window. A peak in the sense strand represents the beginning of a nucleosome while a peak in the antisense strand represents the end. After smoothing, we saw that for most plasma DNA samples, at least 3 well-positioned

nucleosomes downstream of transcription start sites could be detected, and in some cases, up to 5 well-positioned nucleosomes could be detected, in rough accordance to the results of Schones et al. (27) (**Figure 4**). We applied the same analysis on sequence tags of randomly sheared genomic DNA and observed no obvious pattern in tag localization, although the density of tags was higher at the transcription start site (**Figure 4**).

Correction for sequencing bias

Shown in **Figures 10** and **12** are results which may be obtained when sequence tag numbers are treated statistically based on data from the reference human genome. That is, for example, sequence tags from fragments with higher GC content may be overrepresented, and suggest an aneuploidy where none exists. The sequence tag information itself may not be informative, since only a small portion of the fragment ordinarily will be sequenced, while it is the overall G/C content of the fragment that causes the bias. Thus there is provided a method, described in detail in Examples 8 and 10, for correcting for this bias, and this method may facilitate analysis of samples which otherwise would not produce statistically significant results. This method, for correcting for G/C bias of sequence reads from massively parallel sequencing of a genome, comprises the step of dividing the genome into a number of windows within each chromosome and calculating the G/C content of each window. These windows need not be the same as the windows used for calculating sequence tag density; they may be on the order of 10kb-30kb in length, for example. One then calculates the relationship between sequence coverage and G/C content of each window by determining a number of reads per a given window and a G/C content of that window. The G/C content of each window is known from the human genome reference sequence. Certain windows will be ignored, i.e., with no reads or no G/C content. One then assigns a weight to the number of reads per a given window (i.e., the number of sequence tags assigned to that window) based on G/C content, where the weight has a relationship to G/C content such that increasing numbers of reads with increasing G/C content results in decreasing weight per increasing G/C content.

EXAMPLES

The examples below describe the direct sequencing of cell-free DNA from plasma of pregnant women with high throughput shotgun sequencing technology, obtaining on average 5 million sequence tags per patient sample. The sequences obtained were mapped to specific chromosomal locations. This enabled us to measure the over- and under-representation of

chromosomes from an aneuploid fetus. The sequencing approach is polymorphism-independent and therefore universally applicable for the non-invasive detection of fetal aneuploidy. Using this method we successfully identified all 9 cases of trisomy 21 (Down syndrome), 2 cases of trisomy 18 and 1 case of trisomy 13 in a cohort of 18 normal and aneuploid pregnancies; trisomy was detected at gestational ages as early as the 14th week. Direct sequencing also allowed us to study the characteristics of cell-free plasma DNA, and we found evidence that this DNA is enriched for sequences from nucleosomes.

EXAMPLE 1: Subject Enrollment

The study was approved by the Institutional Review Board of Stanford University. Pregnant women at risk for fetal aneuploidy were recruited at the Lucile Packard Children Hospital Perinatal Diagnostic Center of Stanford University during the period of April 2007 to May 2008. Informed consent was obtained from each participant prior to the blood draw. Blood was collected 15 to 30 minutes after amniocentesis or chorionic villus sampling except for 1 sample that was collected during the third trimester. Karyotype analysis was performed via amniocentesis or chorionic villus sampling to confirm fetal karyotype. 9 trisomy 21 (T21), 2 trisomy 18 (T18), 1 trisomy 13 (T13) and 6 normal singleton pregnancies were included in this study. The gestational age of the subjects at the time of blood draw ranged from 10 to 35 weeks (Table 1). Blood sample from a male donor was obtained from the Stanford Blood Center.

EXAMPLE 2: Sample Processing and DNA Quantification

7 to 15ml of peripheral blood drawn from each subject and donor was collected in EDTA tubes. Blood was centrifuged at 1600g for 10 minutes. Plasma was transferred to microcentrifuge tubes and centrifuged at 16000g for 10 minutes to remove residual cells. The two centrifugation steps were performed within 24 hours after blood collection. Cell-free plasma was stored at -80C until further processing and was frozen and thawed only once before DNA extraction. DNA was extracted from cell-free plasma using QIAamp DNA Micro Kit (Qiagen) or NucleoSpin Plasma Kit (Macherey-Nagel) according to manufacturers' instructions. Genomic DNA was extracted from 200µl whole blood of the donors using QIAamp DNA Blood Mini Kit (Qiagen). Microfluidic digital PCR (Fluidigm) was used to quantify the amount of total and fetal DNA using Taqman assays targeting at the EIF2C1 locus on chromosome 1 (Forward: 5' GTTCGGCTTTCACCAGTCT 3' (SEQ ID NO: 1) ; Reverse: 5' CTCCATAGCTCTCCCCACTC 3' (SEQ ID NO: 2); Probe: 5' HEX-

GCCCTGCCATGTGGAAGAT-BHQ1 3' (SEQ ID NO: 3); amplicon size: 81bp) and the SRY locus on chromosome Y (Forward: 5' CGCTTAACATAGCAGAAGCA 3'(SEQ ID NO: 4); Reverse: 5' AGTTTCGAACTCTGGCACCT 3'(SEQ ID NO: 5); Probe: 5' FAM-TGTCGCACTCTCCTTGTTTTTGACA-BHQ1 3'(SEQ ID NO: 6); amplicon size: 84bp) respectively. A Taqman assay targeting at DYS14 (Forward: 5' ATCGTCCATTTCCAGAATCA 3'(SEQ ID NO: 6); Reverse: 5' GTTGACAGCCGTGGAATC 3' (SEQ ID NO: 7); Probe: 5' FAM-TGCCACAGACTGAACTGAATGATTTTC-BHQ1 3' (SEQ ID NO: 8); amplicon size: 84bp), a multi-copy locus on chromosome Y, was used for the initial determination of fetal sex from cell-free plasma DNA with traditional real-time PCR. PCR reactions were performed with 1x iQ Supermix (Bio-Rad), 0.1% Tween-20 (microfluidic digital PCR only), 300nM primers, and 150nM probes. The PCR thermal cycling protocol was 95C for 10 min, followed by 40 cycles of 95C for 15s and 60C for 1 min. Primers and probes were purchased form IDT.

EXAMPLE 3: Sequencing

A total of 19 cell-free plasma DNA samples, including 18 from pregnant women and 1 from a male blood donor, and genomic DNA sample from whole blood of the same male donor, were sequenced on the Solexa/Illumina platform. ~1 to 8ng of DNA fragments extracted from 1.3 to 5.6ml cell-free plasma was used for sequencing library preparation (Table 1). Library preparation was carried out according to manufacturer's protocol with slight modifications. Because cell-free plasma DNA was fragmented in nature, no further fragmentation by nebulization or sonication was done on plasma DNA samples.

Genomic DNA from male donor's whole blood was sonicated (Misonix XL-2020) (24 cycles of 30s sonication and 90s pause), yielding fragments with size between 50 and 400bp, with a peak at 150bp. ~2ng of the sonicated genomic DNA was used for library preparation. Briefly, DNA samples were blunt ended and ligated to universal adaptors. The amount of adaptors used for ligation was 500 times less than written on the manufacturer's protocol. 18 cycles of PCR were performed to enrich for fragments with adaptors using primers complementary to the adaptors. The size distributions of the sequencing libraries were analyzed with DNA 1000 Kit on the 2100 Bioanalyzer (Agilent) and quantified with microfluidic digital PCR (Fluidigm). The libraries were then sequenced using the Solexa 1G Genome Analyzer according to manufacturer's instructions.

Cell-free plasma DNA from a pregnant woman carrying a normal male fetus was also sequenced on the 454/Roche platform. Fragments of DNA extracted from 5.6ml of cell-free plasma (equivalent to ~4.9ng of DNA) were used for sequencing library preparation. The sequencing library was prepared according to manufacturer's protocol, except that no nebulization was performed on the sample and quantification was done with microfluidic digital PCR instead of capillary electrophoresis. The library was then sequenced on the 454 Genome Sequencer FLX System according to manufacturer's instructions.

Electropherograms of Solexa sequencing libraries were prepared from cell-free plasma DNA obtained from 18 pregnant women and 1 male donor. Solexa library prepared from sonicated whole blood genomic DNA from the male donor was also examined. For libraries prepared from cell-free DNA, all had peaks at average 261bp (range: 256-264bp). The actual peak size of DNA fragments in plasma DNA is ~168bp (after removal of Solexa universal adaptor (92bp)). This corresponds to the size of a chromatosome.

EXAMPLE 4: Data Analysis

Shotgun Sequence Analysis

Solexa sequencing produced 36 to 50bp reads. The first 25bp of each read was mapped to the human genome build 36 (hg18) using ELAND from the Solexa data analysis pipeline. The reads that were uniquely mapped to the human genome having at most 1 mismatch were retained for analysis. To compare the coverage of the different chromosomes, a sliding window of 50kb was applied across each chromosome, except in regions of assembly gaps and microsatellites, and the number of sequence tags falling within each window was counted and the median value was chosen to be the representative of the chromosome. Because the total number of sequence tags for each sample was different, for each sample, we normalized the sequence tag density of each chromosome (except chromosome Y) to the median sequence tag density among autosomes. The normalized values were used for comparison among samples in subsequent analysis. We estimated fetal DNA fraction from chromosome 21 for T21 cases, chromosome 18 from T18 cases, chromosome 13 from T13 case, and chromosomes X and Y for male pregnancies. For chromosome 21,18, and 13, fetal DNA fraction was estimated as 2*(x-1), where x was the ratio of the over-represented chromosome sequence tag density of each trisomy case to the median chromosome sequence tag density of the all disomy cases. For chromosome X, fetal DNA was estimated as 2*(1-x), where x was the ratio of chromosome X sequence tag density

of each male pregnancy to the median chromosome X sequence tag density of all female pregnancies. For chromosome Y, fetal DNA fraction was estimated as the ratio of chromosome Y sequence tag density of each male pregnancy to that of male donor plasma DNA. Because a small number of chromosome Y sequences were detected in female pregnancies, we only considered sequence tags falling within transcribed regions on chromosome Y and subtracted the median number of tags in female pregnancies from all samples; this amounted to a correction of a few percent. The width of 99% confidence intervals was calculated for all disomy 21 pregnancies as t*s/vN, where N is the number of disomy 21 pregnancies, t is the t-statistic corresponding to a=0.005 with degree of freedom equals N-1, and s is the standard deviation. A confidence interval gives an estimated range of values, which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. (Definition taken from Valerie J. Easton and John H. McColl's Statistics Glossary v1.1)

To investigate the distribution of sequence tags around transcription start sites, a sliding window of 5bp was applied from -1000bp to +1000bp of transcription start sites of all RefSeq genes on all chromosomes except chromosome Y. The number of sequence tags mapped to the sense and antisense strands within each window was counted. Moving average with a window of 10 data points was used to smooth the data. All analyses were done with Matlab.

We selected the sequence tags that mapped uniquely to the human genome with at most 1 mismatch (on average ~5 million) for analysis. The distribution of reads across each chromosome was examined. Because the distribution of sequence tags across each chromosome was non-uniform (possibly technical artifacts), we divided the length of each chromosome into non-overlapping sliding window with a fixed width (in this particular analysis, a 50kbp window was used), skipping regions of genome assembly gaps and regions with known microsatellite repeats. The width of the window is should be large enough such that there are a sufficient number of sequence tags in each window, and should be small enough such that there are sufficient number of windows to form a distribution. With increasing sequencing depth (i.e., increasing total number of sequence tags), the window width can be reduced. The number of sequence tags in each window was counted. The distribution of the number of sequence tags per 50kb for each chromosome was examined. The median value of the number of sequence tags per 50kb (or 'sequence tag density') for

each chromosome was chosen in order to suppress the effects of any under- or over-represented regions within the chromosome. Because the total number of sequence tags obtained for each sample was different, in order to compare among samples, we normalized each chromosomal sequence tag density value (except chromosome Y) by the median sequence tag density among all autosomes (non-sex chromosomes).

For the 454/Roche data, reads were aligned to the human genome build 36 (hg18, see hyper text transfer protocol (http) genome.ucsc.edu/cgi-bin/hgGateway) using the 454 Reference Mapper. Reads having accuracy of greater than or equal to 90% and coverage (i.e., fraction of read mapped) greater than or equal to 90% were retained for analysis. To study the size distribution of total and fetal DNA, the number of retained reads falling within each 10bp window between 50bp to 330bp was counted. The number of reads falling within different size ranges may be studied, i.e., reads of between 50-60 bp, 60-70 bp, 70-80 bp, etc., up to about 320-330 bp, which is around the maximum read length obtained.

EXAMPLE 5: Genome Data Retrieval

Information regarding G/C content, location of transcription start sites of RefSeq genes, location of assembly gaps and microsatellites were obtained from the UCSC Genome Browser.

EXAMPLE 6 Nucleosome Enrichment

The distribution of sequence tags around transcription start sites (TSS) of RefSeq genes were analyzed (data not shown). The plots were similar to **Figure 4**. Each plot represented the distribution for each plasma DNA or gDNA sample. Data are obtained from three different sequencing runs (P1, P6, P52, P53, P26, P40, P42 were sequenced together; male genomic DNA, male plasma DNA, P2, P7, P14, P19, P31 were sequenced together; P17, P20, P23, P57, P59, P64 were sequenced together). The second batch of samples suffers greater G/C bias as observed from inter- and intra-chromosomal variation. Their distributions around TSS have similar trends with more tags at the TSS. Such trend is not as prominent as in the distributions of samples sequenced in other runs. Nonetheless, at least 3 well-positioned nucleosomes were detectable downstream of transcription start sites for most plasma DNA samples, suggesting that cell-free plasma DNA shares features of nucleosomal DNA, a piece of evidence that this DNA is of apoptotic origin.

EXAMPLE 7: Calculating fetal DNA fraction in maternal plasma of male pregnancies:

i. With Digital PCR Taqman Assays

Digital PCR is the amplification of single DNA molecule. DNA sample is diluted and distributed across multiple compartments such that on average there is less than 1 copy of DNA per compartment. A compartment displaying fluorescence at the end of a PCR represents the presence of at least one DNA molecule.

Assay for Total DNA: EIF2C1 (Chromosome 1)

Assay for Fetal DNA: SRY (Chromosome Y)

The count of positive compartments from the microfluidic digital PCR chip of each assay is converted to the most probable count according to the method described in the supporting information of the following reference: Warren L, Bryder D, Weissman IL, Quake SR (2006) Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. Proc Nat Acad Sci, 103: 17807-12.

Fetal DNA Fraction ε = (SRY count) / (EIF2C1 count / 2)

ii. With Sequence Tags

From ChrX:

Let fetal DNA fraction be ε

	Maternal Contribution	Male Fetus Contribution	Female Fetus Contribution
ChrX	$2(1-\varepsilon)$	ε	2ε

Male pregnancies ChrX sequence tag density (fetal and maternal) = $2(1-\varepsilon) + \varepsilon = 2 - \varepsilon$

Female pregnancies ChrX sequence tag density (fetal and maternal) = $2(1- \varepsilon) + 2\varepsilon = 2$

Let x be the ratio of ChrX sequence tag density of male to female pregnancies. In this study, the denominator of this ratio is taken to be the median sequence tag density of all female pregnancies.

Thus, fetal DNA fraction $\varepsilon = 2(1-x)$

From ChrY:

Fetal DNA fraction ε = (sequence tag density of ChrY in maternal plasma/sequence tag density of ChrY in male plasma)

Note that in these derivations, we assume that the total number of sequence tags obtained is the same for all samples. In reality, the total number of sequence tags obtained for different sample is different, and we have taken into account such differences in our estimation of fetal DNA fraction by normalizing the sequence tag density of each chromosome to the median of the autosomal sequence tag densities for each sample.

Calculating fetal DNA fraction in maternal plasma of aneuploid (trisomy) pregnancies:

Let fetal DNA fraction be ε

	Maternal Contribution	Trisomic Fetus Contribution	Disomic Fetus Contribution
Trisomic Chromosome	2(1-ε)	3ε	2ε

Trisomic pregnancies trisomic chromosome sequence counts (fetal and maternal)

$= 2(1-\varepsilon) + 3\varepsilon = 2 + \varepsilon$

Disomic pregnancies trisomic chromosome sequence counts (fetal and maternal)

$= 2(1- \varepsilon) + 2\ \varepsilon = 2$

Let x be the ratio of trisomic chromosome sequence counts (or sequence tag density) of trisomic to disomic pregnancies. In this study, the denominator of this ratio is taken to be the median sequence tag density of all disomic pregnancies.

Thus, fetal DNA fraction $\varepsilon = 2(x-1)$.

EXAMPLE 8: Correction of sequence tag density bias resulting from G/C or A/T content among different chromosomes in a sample

This example shows a refinement of results indicating sequences mapping to different chromosomes and permitting the determination of the count of different chromosomes or regions thereof. That is, the results as shown in **Figure 1A** may be corrected to eliminate the variations in sequence tag density shown for chromosomes higher in G/C content, shown towards the right of the Figure. This spread of values results from sequencing bias in the method used, where a greater number of reads tend to be obtained depending on G/C content. The results of the method of this example are shown in **Figure 10**. **Figure 10** is an overlay

which shows the results from a number of different samples, as indicated in the legend. The sequence tag density values in **Figs 1** and **10** were normalized to those of a male genomic DNA control, since the density values are not always 1 for all the chromosomes (even after GC correction) but are consistent among a sample. For example, after GC correction, values from all samples for chr19 cluster around 0.8 (not shown). Adjusting the data to a nominal value of 1 can be done by plotting the value relative to the male gDNA control. This makes the values for all chromosomes cluster around 1

Outlying chromosome sequence tag densities can be seen as significantly above a median sequence tag density; disomic chromosomes are clustered about a line running along a density value of about 1. As can be seen there, the results from chromosome 19 (far right, highest in G/C content), for example, show a similar value when disomic as other disomic chromosomes. The variations between chromosomes with low and high G/C content are eliminated from the data to be examined. Samples (such as P13 in the present study) which could not have been unambiguously interpreted now may be. Since G/C content is the opposite of A/T content, the present method will correct for both. Either G/C bias or A/T bias can result from different sequencing methods. For example, it has been reported by others that the Solexa method results in a higher number of reads from sequences where the G/C content is high. See, Dohm et al., "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing," *Nuc. Acids Res.* 36(16), e105; doi:10.1093/nar/gkn425. The procedure of the present example follows the following steps:

a. Calculate G/C content of the human genome. Calculate the G/C content of every 20kb non-overlapping window of each chromosome of the human genome (HG18) using the hgG/CPercent script of the UCSC Genome Browser's "kent source tree," which contains different utility programs, available to the public under license. The output file contains the coordinate of each 20kb bin and the corresponding G/C content. It was found that a large number of reads were obtained higher G/C ranges (about 55-70%) and very few reads were obtained at lower G/C content percentages, with essentially none below about 30% G/C (data not shown). Because the actual length of a sequenced DNA fragment is not known (we only sequenced the first 25bp of one end of a piece of DNA on the flow cell), and it's the G/C content of the entire piece of DNA that contributed to sequencing bias, an arbitrary window of known human genomic DNA sequence is chosen for determining G/C content of different reads. We chose a 20kb window to look at the relationship between

number of reads and GC content. The window can be much smaller e.g., 10kb or 5kb, but a size of 20kb makes computation easier.

b. Calculate the relationship between sequence coverage and G/C content. Assign weight to each read according to G/C content. For each sample, the number of read per 20kb bin is counted. The number of read is plotted against G/C content. The average number of read is calculated for every 0.1% G/C content, ignoring bins with no reads, bins with zero G/C percent, and bins with over-abundant reads. The reciprocal of the average number of reads for a particular G/C percent relative to the global median number of read is calculated as the weight. Each read is then assigned a weight depending on the G/C percent of the 20kb window it falls into.

c. Investigate the distribution of reads across each autosome and chromosome X. In this step, the number of reads, both unweighted and weighted, in each non-overlapping 50kb window is recorded. For counting, we chose a 50kb window in order to obtain a reasonable number of reads per window and reasonable number of windows per chromosome to look at the distributions. Window size may be selected based on the number of reads obtained in a given experiment, and may vary over a wide range. For example, 30K-100K may be used. Known microsatellite regions are ignored. A graph showing the results of chr1 of P7 is shown in **Figure 11**, which illustrates the weight distribution of this step (c) from sample P7, where the weight assigned to different G/C contents is shown; Reads with higher G/C content are overly represented than average and thus are given less weight.

d. Investigate the distribution of reads across chrY. Calculate the number of chrY reads in transcribed regions after applying weight to reads on chrY. Chromosome Y is treated individually because it is short and has many repeats. Even female genome sequence data will map in some part to chromosome Y, due to sequencing and alignment errors. The number of chrY reads in transcribed regions after applying weight to reads on chrY is used to calculate percentage of fetal DNA in the sample.

EXAMPLE 9: Comparing different patient samples using statistical analyses (t statistic)

This example shows another refinement of results as obtained using the previous examples. In this case, multiple patient samples are analyzed in a single process. **Figure 12** illustrates the results of an analysis of patients P13, P19, P31, P23, P26, P40, P42, P1, P2, P6, P7, P14, P17, P20, P52, P53, P57, P59 and P64, with their respective karyotypes indicated, as

in **Table 1**, above. The dotted line shows the 99% confidence interval, and outliers may be quickly identified. It may be seen by looking below the line that male fetuses have less chromosome X (solid triangles). An exception is P19, where it is believed that there were not enough total reads for this analysis. It may be seen by looking above the line that trisomy 21 patients (solid circles) are P 1, 2, 6, 7, 14, 17, 20, 52 and 53. P57 and 59 have trisomy 18 (open diamonds) and P64 has trisomy 13 (star). This method may be presented by the following three step process:

Step 1: Calculate a t statistic for each chromosome relative to all other chromosome in a sample. Each t statistic tells the value of each chromosome median relative to other chromosomes, taking into account the number of reads mapped to each chromosome (since the variation of the median scales with the number of reads). As described above, the present analyses yielded about 5 million reads per sample. Although one may obtain 3-10 million reads per sample, these are short reads, typically only about 20-100 bp, so one has actually only sequenced, for example about 300 million of the 3 billion bp in the human genome. Thus, statistical methods are used where one has a small sample and the standard deviation of the population (3 billion, or 47 million for chromosome 21) is unknown and it is desired to estimate it from the sample number of reads in order to determine the significance of a numerical variation. One way to do this is by calculating Student's t-distribution, which may be used in place of a normal distribution expected from a larger sample. The t-statistic is the value obtained when the t-distribution is calculated. The formula used for this calculation is given below. Using the methods presented here, other t-tests can be used.

Step 2: Calculate the average t statistic matrix by averaging the values from all samples with disomic chromosomes. Each patient sample data is placed in a t matrix, where the row is chr1 to chr22, and the column is also chr1 to chr22. Each cell represents the t value when comparing the chromosomes in the corresponding row and column (i.e., position (2,1) in the matrix is the t-value of when testing chr2 and chr1) the diagonal of the matrix is 0 and the matrix is symmetric. The number of reads mapping to a chromosome is compared individually to each of chr1-22.

Step 3: Subtract the average t statistic matrix from the t statistic matrix of each patient sample. For each chromosome, the median of the difference in t statistic is selected as the representative value.

The t statistic for 99% confidence for large number of samples is 3.09. Any chromosome with a representative t statistic outside -3.09 to 3.09 is determined as non-disomic.

EXAMPLE 10: Calculation of required number of sequence reads after G/C bias correction

In this example, a method is presented that was used to calculate the minimum concentration of fetal DNA in a sample that would be needed to detect an aneuploidy, based on a certain number of reads obtained for that chromosome (except chromosome Y). **Figure 13** and **Figure 14** show results obtained from 19 patient plasma DNA samples, 1 donor plasma DNA sample, and duplicate runs of a donor gDNA sample. It is estimated in **Figure 13** that the minimum fetal DNA % of which over-representation of chr21 can be detected at the best sampling rate (~70k reads mapped to chr21) is ~6%. (indicated by solid lines in **Fig. 13**). The lines are drawn between about 0.7 $X10^5$ reads and 6% fetal DNA concentration. It can be expected that higher numbers of reads (not exemplified here) the needed fetal DNA percentage will drop, probably to about 4%.

In **Figure 14**, the data from **Figure 13** are presented in a logarithmic scale. This shows that the minimum required fetal DNA concentration scales linearly with the number of reads in a square root relationship (slope of -.5). These calculations were carried out as follows:

For large n (n>30), t statistic $t = \frac{\overline{y_2} - \overline{y_1}}{\sqrt{\frac{{s_2}^2}{n_2} + \frac{{s_1}^2}{n_1}}}$, where $\overline{y_2} - \overline{y_1}$ is the difference in means (or amount of over- or under-representation of a particular chromosome) to be measured; s is the standard deviation of the number of reads per 50kb in a particular chromosome; n is the number of samples (i.e., the number of 50kb windows per chromosome). Since the number of 50kb windows per chromosome is fixed, $n_1 = n_2$. If we assume that $s_1 \approx s_2$, $\overline{y_2} - \overline{y_1} \approx t\sqrt{\frac{2{s_1}^2}{n_1}}$ = sqrt(2)*half width of the confidence interval at confidence level governed by the value of t. Thus, $\frac{\overline{y_2}}{\overline{y_1}} - 1 \approx \frac{t\sqrt{\frac{2{s_1}^2}{n_1}}}{\overline{y_1}}$. For every chromosome

in every sample, we can calculate the value $\frac{t\sqrt{\frac{2s_1^2}{n_1}}}{\overline{y_1}}$, which corresponds to the minimum over- or under-representation ($\frac{\overline{y_2}}{\overline{y_1}}-1$) that can be resolved with confidence level governed by the value of t. Note that $2*(\frac{\overline{y_2}}{\overline{y_1}}-1)*100\%$ corresponds to the minimum fetal DNA % of which any over- or under-representation of chromosomes can be detected. We expect the number of reads mapped to each chromosome to play a role in determining standard deviation s_1, since according to Poisson distribution, the standard deviation equals to the square root of the mean. By plotting $2*(\frac{\overline{y_2}}{\overline{y_1}}-1)*100\%$ vs. number of reads mapped to each chromosome in all the samples, we can evaluate the minimum fetal DNA % of which any over- or under-representation of chromosomes can be detected given the current sampling rate.

After correction of G/C bias, the number of reads per 50kb window for all chromosomes (except chromosome Y) is normally distributed. However, we observed outliers in some chromosomes (e.g., a sub-region in chromosome 9 has near zero representation; a sub-region in chromosome 20 near the centromere has unusually high representation) that affect the calculation of standard deviation and the mean. We therefore chose to calculate confidence interval of the median instead of the mean to avoid the effect of outliers in the calculation of confidence interval. We do not expect the confidence interval of the median and the mean to be very different if the small number of outliers has been removed. The 99.9% confidence interval of the median for each chromosome is estimated from bootstrapping 5000 samples from the 50kb read distribution data using the percentile method. The half width of the confidence interval is estimated as 0.5*confidence interval. We plot 2*(half width of confidence interval of median)/median*100% vs. number of reads mapped to each chromosome for all samples.

Bootstrap resampling and other computer-implemented calculations described here were carried out in MATLAB®, available from The Mathworks, Natick, MA.

CONCLUSION

The above specific description is meant to exemplify and illustrate the invention and should not be seen as limiting the scope of the invention, which is defined by the literal and equivalent scope of the appended claims. Any patents or publications mentioned in this specification are intended to convey details of methods and materials useful in carrying out certain aspects of the invention which may not be explicitly set out but which would be understood by workers in the field. Such patents or publications are hereby incorporated by reference to the same extent as if each was specifically and individually incorporated by reference, as needed for the purpose of describing and enabling the method or material referred to.

REFERENCES

1. Cunningham F, et al. (2002) in Williams Obstetrics (McGraw-Hill Professional, New York), p. 942.

2. (2007) ACOG Practice Bulletin No. 88, December 2007. Invasive prenatal testing for aneuploidy. *Obstet Gynecol*, 110: 1459-1467.

3. Wapner R, et al. (2003) First-trimester screening for trisomies 21 and 18. *N Engl J Med*, 349: 1405-1413.

4. Alfirevic Z, Neilson JP (2004) Antenatal screening for Down's syndrome. *Bmj* 329: 811-812.

5. Malone FD, et al. (2005) First-trimester or second-trimester screening, or both, for Down's syndrome. *N Engl J Med*, 353: 2001-2011.

6. Herzenberg LA, et al. (1979) Fetal cells in the blood of pregnant women: detection and enrichment by fluorescence-activated cell sorting. *Proc Natl Acad Sci USA*, 76: 1453-1455.

7. Bianchi DW, et al. (1990) Isolation of fetal DNA from nucleated erythrocytes in maternal blood. *Proc Natl Acad Sci USA*, 87: 3279-3283.

8. Cheung MC, Goldberg JD, Kan YW (1996) Prenatal diagnosis of sickle cell anaemia and thalassaemia by analysis of fetal cells in maternal blood. *Nat Genet*, 14: 264-268.

9. Bianchi DW, et al. (1997) PCR quantitation of fetal cells in maternal blood in normal and aneuploid pregnancies. *Am J Hum Genet*, 61: 822-829.

10. Bianchi DW, et al. (2002) Fetal gender and aneuploidy detection using fetal cells in maternal blood: analysis of NIFTY I data. National Institute of Child Health and Development Fetal Cell Isolation Study. *Prenat Diagn*, 22: 609-615.

11. Lo YM, et al. (1997) Presence of fetal DNA in maternal plasma and serum. *Lancet*, 350: 485-487.

12. Dennis Lo YM, Chiu RW (2007) Prenatal diagnosis: progress through plasma nucleic acids. *Nat Rev Genet*, 8: 71-77.

13. Lo YM, et al. (1998) Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *Am J Hum Genet*, 62: 768-775.

14. Lo YM, et al. (2007) Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection. *Nat Med*, 13: 218-223.

15. Tong YK, et al. (2006) Noninvasive prenatal detection of fetal trisomy 18 by epigenetic allelic ratio analysis in maternal plasma: Theoretical and empirical considerations. *Clin Chem*, 52: 2194-2202.

16. Dhallan R, et al. (2007) A non-invasive test for prenatal diagnosis based on fetal DNA present in maternal blood: a preliminary study. *Lancet*, 369: 474-481.

17. Fan HC, Quake SR (2007) Detection of aneuploidy with digital polymerase chain reaction. *Anal Chem*, 79: 7576-7579.

18. Lo YM, et al. (2007) Digital PCR for the molecular detection of fetal chromosomal aneuploidy. *Proc Natl Acad Sci USA*, 104: 13116-13121.

19. Quake SR, Fan HC. (2006). Non-invasive fetal genetic screening by digital analysis. USA Provisional Patent Application No. 60/764,420. 20. Mardis ER (2008) Next-Generation DNA Sequencing Methods. *Annu Rev Genomics Hum Genet*, 9: 387-402.

20. Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. *Nature*, 409: 860-921.

21. Chan KC, et al. (2004) Size distributions of maternal and fetal DNA in maternal plasma. *Clin Chem*, 50: 88-92.

22. Li Y, et al. (2004) Size separation of circulatory DNA in maternal plasma permits ready detection of fetal DNA polymorphisms. *Clin Chem*, 50: 1002-1011.

23. Cooper G, Hausman R (2007) in The cell: a molecular approach (Sinauer Associates, Inc, Sunderland), p. 168.

24. Jahr S, et al. (2001) DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells. *Cancer Res*, 61: 1659-1665.

25. Giacona MB, et al. (1998) Cell-free DNA in human blood plasma: length measurements in patients with pancreatic cancer and healthy controls. *Pancreas*, 17: 89-97.

26. Schones DE, et al. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132: 887-898.

27. Ozsolak F, Song JS, Liu XS, Fisher DE (2007) High-throughput mapping of the chromatin structure of human promoters. *Nat Biotechnol*, 25: 244-248.

28. Yuan GC, et al. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, 309: 626-630.

29. Lee W, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet*, 39: 1235-1244.

30. Sohda S, et al. (1997) The proportion of fetal nucleated red blood cells in maternal blood: estimation by FACS analysis. *Prenat Diagn*, 17: 743-752.

31. Hamada H, et al. (1993) Fetal nucleated cells in maternal peripheral blood: frequency and relationship to gestational age. *Hum Genet*, 91: 427-432.

32. Nelson JL (2008) Your cells are my cells. *Sci Am*, 298: 64-71.

33. Khosrotehrani K, Bianchi DW (2003) Fetal cell microchimerism: helpful or harmful to the parous woman? *Curr Opin Obstet Gynecol*, 15: 195-199.

34. Lo YM, et al. (1999) Rapid clearance of fetal DNA from maternal plasma. *Am J Hum Genet*, 64: 218-224.

35. Smid M, et al. (2003) No evidence of fetal DNA persistence in maternal plasma after pregnancy. *Hum Genet*, 112: 617-618.

36. Rijnders RJ, Christiaens GC, Soussan AA, van der Schoot CE (2004) Cell-free fetal DNA is not present in plasma of nonpregnant mothers. *Clin Chem*, 50: 679-681; author reply 681.

37. Hillier LW, et al. (2008) Whole-genome sequencing and variant discovery in C. elegans. *Nat Methods*, 5: 183-188.

38. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*

39. Harris TD, et al. (2008) Single-molecule DNA sequencing of a viral genome. *Science*, 320: 106-109.

40. Samura O, et al. (2003) Cell-free fetal DNA in maternal circulation after amniocentesis. *Clin Chem*, 49: 1193-1195.

41. Lo YM, et al. (1999) Increased fetal DNA concentrations in the plasma of pregnant women carrying fetuses with trisomy 21. *Clin Chem*, 45: 1747-1751.

42. Segal E, et al. (2006) A genomic code for nucleosome positioning. *Nature*, 442: 772-778.

CLAIMS

What is claimed is:

1. A method of testing for an abnormal distribution of a specified chromosome portion in a mixed sample of normally and abnormally distributed chromosome portions obtained from a subject, comprising:

   (a) obtaining sequences from multiple chromosome portions of the mixed sample to obtain a number of sequence tags of sufficient length of determined sequence to be assigned to a chromosome location within a genome;

   (b) assigning the sequence tags to corresponding chromosome portions including at least the specified chromosome by comparing the sequence to a reference genomic sequence;

   (c) determining values for numbers of sequence tags mapping to chromosome portions by using a number of windows of defined length within normally and abnormally distributed chromosome portions to obtain a first value and a second value therefrom; and

   (d) using the values from step (c) to determine a differential, between the first value and the second value, which is determinative of whether or not the abnormal distribution exists.

2. The method of claim 1 wherein to determine a differential includes the step of comparing a normalized sequence tag density of the specified DNA chromosome portion to a normalized sequence tag density of another DNA chromosome portion in said mixed sample, wherein all autosomes are used to calculate the normalized sequence tag density.

3. The method of claim 1 wherein the mixed sample is a mixture of maternal and fetal DNA.

4. The method of claim 1 wherein the mixed sample is a mixture of normal and genetically altered DNA from a tumor.

5. The method of claim 1 wherein the sequencing is massively parallel sequencing.

6. The method of claim 1 wherein the abnormal distribution is an aneuploidy of at least one of chromosome 13, 18 and 21.

7. The method of claim 1 wherein the step of assigning sequence tags to corresponding chromosome portions allows one mismatch.

8. The method of claim 1 wherein the sequence tags are about 25-100 bp in length.

9. The method of claim 8 wherein at least about 1 million sequence tags are obtained.

10. The method of claim 8 further comprising the steps of calculating a normalized sequence tag density of the specified DNA chromosome portion to a normalized sequence tag density of another DNA chromosome portion in said mixed sample.

11. The method of claim 10 wherein the step of calculating a differential includes the step of comparing a normalized sequence tag density of the specified DNA chromosome portion to a normalized sequence tag density of another DNA chromosome portion in said mixed sample, wherein all autosomes are used to calculate the normalized sequence tag density.

12. The method of claim 11 further comprising the step of measuring over- and under-representation of a chromosome by determining a sequence tag density for each chromosome in the sample, namely chromosomes 1-22, X and also chromosome Y if present.

13. A method for measuring a fetal DNA fraction in a sample of mixed maternal and fetal DNA, wherein a chromosome having trisomy is measured by being $(1 + \varepsilon/2)$ of that of a corresponding disomy, where ε represents the fetal DNA fraction.

14. A noninvasive method for enumerating an abnormally distributed specified chromosome or portion thereof in a mixed sample of DNA from the plasma of a pregnant subject, comprising:

(a) sequencing multiple portions of the mixed sample of DNA to obtain at least one million sequence tags of sufficient length of determined sequence to be assigned to a chromosome location within a genome;

(b) determining a number representing a number of sequence tags within a number of discrete windows each of which is a subsequence of a chromosome and is between about 10Kb and 100Kb in length; and

(c) calculating a differential between a number on a normal chromosome and a number on the abnormally distributed chromosome which is determinative of whether or not the abnormally distributed chromosome is present or not.

15. The method of claim 14 wherein the chromosome is any one of X, Y, 18, 21, 17 or 13.

16. The method of claim 14 wherein said calculating a differential comprises obtaining a sequence density of the abnormally distributed chromosome and comparing it to a value of a disomic chromosome.

17. The method of claim 14 wherein said number of discrete windows are comprised of sliding non overlapping windows of 10-100 kb extending along substantially an entire chromosome.

18. The method of claim 14 further comprising the step of measuring a number of sequence tags within transcriptional start sites.

19. A method of determining an abnormally distributed chromosome portion of interest in a mixed sample of normally and abnormally distributed DNA molecules, comprising:

(a) sequencing DNA in said sample by massively parallel sequencing to obtain a number of sequence tags

(b) mapping said sequence tags to specific chromosome portions, each chromosomal portion being comprised in a sliding window of a predetermined length;

(c) determining numbers of sequence tags mapped to each sliding window on at least each autosome;

(d) determining a mean of said numbers for each autosome and a second mean for at least all autosomes;

(e) calculating a normalized value from all autosomes, using said second mean; and

(f) comparing normalized values among autosomes to determine any abnormally distributed autosomal chromosome portion of interest.

20. The method of claim 19 further comprising the step of calculating a relationship between numbers of sequence tags and GC content associated with sequence tags in a given sliding window and correcting for a higher number of reads resulting from a change in GC content.

21. The method of claim 19 further comprising the step of calculating a t statistic for each chromosome relative to other chromosomes in the mixed sample, whereby each t statistic indicates a value of a chromosome relative to other chromosomes in a sample, said value being indicative of disomy.

22. The method of claim 19 further comprising the step of calculating a normalized value for chromosome X and, if present, Y.

23. The method of claim 19 wherein said mapping includes mapping sequences with one mismatch.

24. A method for correcting for GC bias of sequence reads from massively parallel sequencing of a genome, comprising the steps of:

(a)dividing the genome into a number of windows within each chromosome and calculating the GC content of each window;

(b)calculating the relationship between sequence coverage and GC content of each window by determining a number of reads per a given window and a GC content of that window; and

(c)assigning a weight to said number of reads per a given window based on GC content, where the weight has a relationship to GC content such that increasing numbers of reads with increasing GC content results in decreasing weight per increasing GC content.