

FRAGMENTOMICS-BASED PLASMA SEPARATION ASSAYS**FIELD OF THE INVENTION**

This invention is in the field of cfDNA quantification and cfDNA quality control.

BACKGROUND

5 Cell-free DNA (also called circulating free DNA; cfDNA) is DNA released from cells into the blood circulation and into other body fluids. This can be used for numerous clinical applications, notably including the monitoring of tumour-derived cfDNA for cancer diagnostics and therapeutics and the screening of foetal disease using maternal blood. There is an increased interest in the detection and analysis of cfDNA as a powerful, low invasiveness, rapid, cheap diagnostic: for instance, performing
10 a simple blood test to detect cancer (sometimes referred to as a 'liquid biopsy') is highly advantageous relative to other methods of cancer diagnosis such as e.g. a tissue biopsy.

However, a reliable analysis of blood cfDNA is dependent on obtaining a pure sample of plasma through purification from whole blood and applications that are based on cell free DNA, such as cancer diagnostics, rely on high rate of plasma purity. However, because cfDNA extraction methods can easily
15 lead to contamination of the plasma by cellular genomic DNA (originating from ruptured blood cells during the purification process or from contamination of buffy coat during the separation of plasma from the blood), it is important to ascertain plasma purity. Existing methods to evaluate plasma purity tend to be laborious and expensive (and in some cases, insufficiently accurate), such as the use of blood cell counts (e.g., using a hemocytometer or a flow cytometer). More recently, other methods have been
20 explored which exploit the differences in fragment length between cfDNA (mostly fragments of less than 200 or 300 bp in length according to Chan et al., 2004, Clinical Chemistry, 50(1): 88-92) and genomic DNA (typically >10 kbp-long fragments for white blood cells). Gel electrophoresis can be used but requires large amounts of DNA. A fragment analyser may also be used.

Another assay quantitates the relative amounts of short DNA fragments (typically between 70-150 bp)
25 originating from cfDNA as compared to long DNA fragments (typically between 350-600 bp in length) originating from cellular DNA. The relative amount of these fragments can be determined, e.g., using quantitative PCR, as an indication of the level of purity of cfDNA samples (WO 2019/162941). A drawback of this method, however, resides in the requirement to amplify fragments of different sizes. PCR efficiency can vary according to the size of the fragment which is amplified, at least in part
30 because the amplification of longer DNA fragments has been shown to be more sensitive to the presence of PCR inhibitors. Because PCR inhibitors are common in samples originating from blood (e.g., heme, IgG are PCR inhibitors), the co-amplification of fragments of different sizes from blood samples can suffer from impaired amplification of the larger, contaminant, genomic DNA fragment.

There is a need for improved methods and kits for determining the efficiency of plasma separation
35 from whole blood, which are simple to operate, cost-effective and accurate.

DISCLOSURE OF THE INVENTION

The inventors have observed that the fragmentation patterns in cfDNA result in some loci being present at different concentrations than other loci in plasma DNA, but this is not the case for whole blood DNA. Furthermore, these differences in concentration result in differences in amplification patterns for plasma DNA which are not seen in whole blood DNA. The inventors now show that these differences can be reliably exploited in order to determine the efficiency of plasma separation from whole blood.

In one embodiment, a method according to the invention comprises the following steps:

- 10 (a) generating from a plasma DNA sample, by PCR amplification, a first amplification product from a first genomic locus using a first primer pair, wherein the first locus is present at different concentrations in cfDNA and whole blood DNA;
- (b) quantitating the amplification product; and
- 15 (c) determining if the plasma sample is separated based on the quantitated amplification product.

The above embodiment is provided merely to illustrate specific aspects of the invention, which is described in more detail below.

cfDNA fragmentation results in specific patterns of amplification during whole-genome sequencing

It is generally thought that cfDNA is generated by genomic fragmentation during cellular apoptosis and necrosis (Ma *et al.*, 2017, PLoS ONE 12(1): e0169231). This fragmentation has been described to be biased, and as a result, the coverage depth in cfDNA whole-genome sequencing varies across the genome (Ivanov *et al.*, 2015, BMC Genomics 16 (Suppl 13):S1, Ivanov *et al.*, 2019, Front Genet. 10:194.). In other words, some genomic loci are under-represented in plasma DNA as compared to whole blood DNA, but other genomic loci are over-represented. The typical size distribution which has been described in the literature for cfDNA (peaking at about 170bp in human plasma) is consistent with the hypothesis that nucleosomes protect the DNA from fragmentation during apoptosis/necrosis and that nucleosome positions therefore largely determine the boundaries of cfDNA fragments (Ma *et al.*, 2017). Other yet unknown factors possibly also contribute to cfDNA fragmentation patterns that can be exploited in the present invention.

30 The inventors have found that the specific patterns of read depth obtained from whole genome sequencing of plasma DNA are different from those seen for whole blood DNA (see Example 1). Indeed, substantial variation in coverage depth is observed for plasma DNA, but very little variation in coverage depth is observed for the same genomic loci in whole blood DNA. Without wishing to be bound by theory, this variation is thought to arise because plasma DNA is mostly composed of cfDNA

(which is fragmented) and whole blood DNA is mostly composed of genomic DNA (which is less fragmented), and some genomic loci are more concentrated in comparison to whole blood DNA, whereas other genomic loci are less concentrated.

General design of the assay to determine the efficiency of plasma separation

5 The differences in availability of targets for amplification between cfDNA and whole blood DNA can be exploited in methods of the invention. The assay may include amplifying one locus which is present at different concentrations in cfDNA and whole blood DNA. The assay may include amplifying a pair of loci, whereby the concentration ratio of the first genomic locus to the second genomic locus in
10 cfDNA is different from the concentration ratio of the first genomic locus to the second genomic locus in whole blood DNA.

PCR amplification is performed using a plasma DNA sample. The plasma DNA sample may be isolated from plasma using methods known in the art. Extraction of the DNA may be part of the present method, or it may have been performed previously (*e.g.* elsewhere). In other embodiments, PCR is performed directly on a plasma sample, which contains plasma DNA.

15 As used herein, the term "plasma DNA" refers to the DNA from the plasma fraction of whole blood. Plasma DNA contains cfDNA, and it may also contain contaminant cellular genomic DNA. As used herein, the terms "DNA from whole blood" and "whole blood DNA" are used interchangeably and refer to the mixture of DNA obtained from whole blood, *i.e.*, genomic DNA from cells and cfDNA. As used herein, the terms "cell-free DNA" and "circulating cell-free DNA" (abbreviated "cfDNA") are
20 used interchangeably and refer to DNA molecules freely circulating in the blood outside cells. As used herein, the term "genomic DNA" refers to intact DNA molecules in whole blood which are released from the inside of nucleated (white) blood cells.

As used herein, the phrases "determining plasma separation efficiency", "determining if a plasma sample is separated", "determining if a plasma sample is sufficiently/efficiently separated" and the like
25 are interchangeable and refer to determining whether the level of any contaminating white blood cell DNA in the plasma is such that it does not interfere with analysis of the cell-free DNA. Different diagnostic applications involving analysis of cell-free DNA may require different levels of purity of the plasma (*i.e.*, may be characterized by different levels of genomic DNA contamination that are tolerated). Thus, a threshold above which a plasma sample is determined to be separated, such as a
30 threshold Cq of an amplicon, or a ΔCq between two amplified amplicons, according to the present invention, may be different for different assays. The threshold may be set based on the requirements of a particular diagnostic assay or intended use of the plasma.

The assay may include quantifying, from high-throughput sequencing data, the read counts of a set of first genomic loci and the read counts of a set of second genomic loci, wherein the ratio of the sum of
35 the concentrations of the genomic loci of the first set of genomic loci to the sum of the concentrations

of the genomic loci of the second set of genomic loci in cfDNA is different from the ratio of the sum of the concentrations of the genomic loci of the first set of genomic loci to the sum of the concentrations of the genomic loci of the second set of genomic loci in whole blood DNA.

5 The assay may include quantifying, from high-throughput sequencing data of a plasma sample, the read counts of a set of first genomic loci, wherein the sum of the concentrations of the genomic loci of the first set of genomic loci is present at different concentrations in cfDNA and whole blood DNA.

The high-throughput sequencing data are from a plasma DNA sample.

10 The terms "genomic locus" or "locus" as used herein are interchangeable and refer to a DNA sequence at a specific position on a chromosome. The specific position(s) may be identified by the molecular location, namely, by the numbers of the starting and ending base pairs on the chromosome, or, where the locus is a single nucleotide, by its single position. Loci include gene sequences as well as other genomic elements (*e.g.*, intergenic sequences).

Method using PCR

15 The assay for evaluating the quality of plasma separation disclosed herein may include a PCR with primers which produce an amplicon from a first genomic locus present at different concentrations in cfDNA and whole blood DNA.

Selection of loci which distinguish between plasma DNA and whole blood DNA

Method using two loci

20 Preferably, the present invention includes the co-amplification of two loci. Advantageously, and as further explained below, determining if the plasma sample is separated based on two loci involves comparing the *difference* in quantitated amplification products between two loci, thereby freeing the assay from any dependency on the amount of template plasma DNA used for PCR.

25 For the method using two loci, a first and second genomic loci are selected such that the concentration ratio of the first genomic locus to the second genomic locus in cfDNA is different from the concentration ratio of the first genomic locus to the second genomic locus in whole blood DNA. This is possible because the inventors have found that some genomic loci are more concentrated in plasma DNA, and other genomic loci are less concentrated in plasma DNA, relative to whole blood DNA.

30 As used herein, a 'High' locus is enriched in cfDNA relative to whole blood DNA, and a 'Low' locus is depleted in cfDNA relative to whole blood DNA. In other words, a 'High' locus is present at a higher concentration in cfDNA relative to whole blood DNA, and a 'Low' locus is present at a lower concentration in cfDNA relative to whole blood DNA.

When a 'High' locus (*e.g.*, the locus according to SEQ ID NO: 1) is co-amplified with a locus which is not as enriched in cfDNA relative to whole blood DNA as compared to the 'High' locus (*e.g.*, a

'Reference' locus, such as the locus according to SEQ ID NO:3), then the difference in amplification levels between the 'High' locus and the 'Reference' locus is larger for cfDNA than for whole blood DNA. Similarly, when a 'Low' locus (e.g., the locus according to SEQ ID NO: 2) is co-amplified with a locus which is not as depleted in cfDNA relative to whole blood DNA as compared to the 'Low' locus (e.g., a 'Reference' locus, such as the locus according to SEQ ID NO:3), then the difference in amplification levels between the 'Low' locus and the 'Reference' locus is larger for cfDNA than for whole blood DNA.

A 'Reference' locus has a concentration ratio in cfDNA relative to whole blood that is different from the concentration ratio in cfDNA relative to whole blood of the 'High' or 'Low' locus which it is co-amplified with.

In some embodiments, the method according to the invention comprises: (a) generating from a plasma DNA sample, by PCR amplification, a first amplification product from a first genomic locus using a first primer pair, and a second amplification product from a second genomic locus using a second primer pair, wherein the concentration ratio of the first genomic locus to the second genomic locus in cfDNA is different from the concentration ratio of the first genomic locus to the second genomic locus in whole blood DNA; (b) quantitating the amplification products; and (c) determining if the plasma sample is separated based on the quantitated amplification products.

In some embodiments, step (b) includes determining a quantification cycle (Cq) for the first and second amplification products, and wherein the plasma sample is determined to be separated based on a comparison of the difference in Cq for the first and second amplification product (ΔCq) to a predefined threshold.

In some embodiments, the difference between the concentration ratios of the first and second loci in cfDNA and the concentration ratios of the first and second loci in whole blood DNA is at least 1.5-fold. In some embodiments, the difference between the concentration ratios of the first and second loci in cfDNA and the concentration ratios of the first and second loci in whole blood DNA is at least 2-fold. In some embodiments, the difference between the concentration ratios of the first and second loci in cfDNA and the concentration ratios of the first and second loci in whole blood DNA is at least 3-fold. In some embodiments, the difference between the concentration ratios of the first and second loci in cfDNA and the concentration ratios of the first and second loci in whole blood DNA is at least 4-fold. In some embodiments, the difference between the concentration ratios of the first and second loci in cfDNA and the concentration ratios of the first and second loci in whole blood DNA is more than 2-fold.

Selecting suitable genomic loci can be performed, for example, by 2 or 3 steps as follows:

1. Selecting two genomic loci, whereby the concentration ratio of the first genomic locus to the second genomic locus in cfDNA is different from the concentration ratio of the first genomic locus to the second genomic locus in whole blood DNA.

This may be achieved, for example, by selecting two loci which have different read counts within a whole genome sequencing trace of (uncontaminated) plasma DNA that plots read count (i.e., sequencing depth) as a function of the genomic locus. For example, it may be achieved by selecting a 'peak' region to be the first genomic locus and a 'trough' region to be the second genomic locus in a whole genome sequencing trace of plasma DNA. Typically, the bigger the difference in read count between the first and second genomic loci in plasma DNA, the better the pair of loci is suitable for the assay of the present invention. Selecting a pair of loci on the basis of a whole genome sequencing trace of plasma DNA alone relies on the assumption that read count is less variable across genomic loci for a whole blood DNA trace than for a plasma DNA trace. Based on that assumption, choosing two loci which differ in read count in plasma DNA should result in obtaining a first and second locus whereby the concentration ratio of the first genomic locus to the second genomic locus in cfDNA is different from the concentration ratio of the first genomic locus to the second genomic locus in whole blood DNA.

Alternatively, pairs of loci may be selected by comparing a whole genome sequencing trace that plots read count (i.e., sequencing depth) as a function of the genomic locus for plasma DNA to another whole genome sequencing trace that plots read count (i.e., sequencing depth) as a function of the genomic locus for whole blood DNA. Potentially suitable pairs of loci will include a first locus and second locus for which the ratio of the read count between the first locus and the second locus in plasma DNA is different from the ratio of the read count between the first locus and the second locus in whole blood DNA.

For example, a suitable pair of loci may be first and second loci for which the first locus has a different read count to the second locus in plasma DNA but not in whole blood DNA. For example, a suitable pair of loci may be first and second loci for which the first locus which has a different read count to the second locus in whole blood DNA but not in plasma DNA. For example, a suitable pair of loci may consist of a first locus and a second locus, wherein the first locus has a different read count to the second locus in both whole blood DNA and plasma DNA, provided that the ratio of the read count between the first locus and the second locus in plasma DNA is different from the ratio of the read count between the first locus and the second locus in whole blood DNA. Typically, the bigger the difference in relative read count between the first and second genomic loci in plasma DNA and whole blood DNA, the better the pair of loci is suitable for the assay of the present invention.

One way of expressing coverage of a particular site (i.e., a single nucleotide position) from whole genome sequencing data is using 'HitspanN', such as 'Hitspan100'. HitspanN (where N is a positive

even integer, such as 100) refers to the number of sequence reads which span the site with at least $N/2$ nucleotides both upstream and downstream. For example, Hitspan100 refers to the number of sequence reads which span the site with at least 50 nucleotides both upstream and downstream. For example, a Hitspan100 of 90 at a specific site means that there are 90 sequence reads which span this site with at least 50 nucleotides both upstream and downstream. Figure 1A shows a plot of normalised Hitspan100 values. Advantageously, the use of HitspanN (*e.g.* where N is 100) is more successful in visualising loci that are useful to the present invention relative to using sequencing depth. This is because a high Hitspan100 is achieved for a plasma sample only if there is very little or no fragmentation of cfDNA across the sequence which spans a site with at least 50 nucleotides both upstream and downstream. Because methods of the invention typically amplify a locus which is about 100bp in length, the advantage of using Hitspan100 lies in its ability to identify sites which remain highly intact across 50bp either side of the site or on the other hand, are fragmented 50bp either side of the site.

Nucleosome positioning data can also be used for the selection of loci. Loci at nucleosome sites are more likely to be 'High' loci (*i.e.*, enriched in plasma DNA relative to whole blood DNA) and loci between nucleosomes are more likely to be 'Low' loci (*i.e.*, depleted in plasma DNA relative to whole blood DNA). The skilled person will understand that any data which might indicate variations in concentration of loci within plasma DNA, and/or between plasma DNA and whole blood DNA, can be exploited to select genomic loci that may be suitable for use in the present invention.

The loci are generally selected such that the lengths (in nucleotides) of the first and second amplification products are within 35% of their mean length. Preferably, the lengths (in nucleotides) of the first and second amplification products are the same or are within 15% of their mean length.

As an example, if the first amplification product has a length of 100 nucleotides and the second amplification product has a length of 150 nucleotides, then their mean length is given by $(100+150)/2=125$ nucleotides. In this case, the lengths of the first and second amplification products fall within 35% of the mean length because they fall within the range of 81.25 nucleotides ($125-(0.35*125)$) to 168.75 nucleotides ($125+(0.35*125)$). However, the lengths of the first and second amplification products do not fall within 15% of the mean length because they are outside of the range of 106.25 to 143.75 nucleotides.

Co-amplifying fragments of similar lengths is advantageous because co-amplification of fragments of different sizes typically suffers from unequal PCR efficiency. As a result, the relative numbers of amplicons obtained from co-amplification of a short locus and a long locus may not necessarily be a good reflection of the relative concentrations of the short and long locus contained in the original sample. However, the present method aims to obtain a measure of the relative concentrations of the first and second genomic loci in a sample. While the present method can tolerate some differences between the PCR efficiency of generating the first amplification product from the first genomic locus

and the PCR efficiency of generating the second amplification product from the second genomic locus, it is preferable to design the genomic loci with similar lengths so as to minimise any resulting differences in PCR efficiencies. This can be achieved by co-amplifying fragments of similar lengths, such as amplicons which are within 35% (or, preferably, within 15%) of their mean length.

- 5 Alternatively, the loci can be selected such that the first and second amplification products are <350 nucleotides in length. PCR inhibitors are common in samples originating from blood such as plasma samples (e.g., heme and IgG are PCR inhibitors). The amplification of longer DNA fragments has been shown to be more sensitive to the presence of PCR inhibitors relative to the amplification of shorter DNA fragments. Therefore, it is advantageous to select shorter loci for the present method, e.g., such
10 that the first and second amplification products are both <350 nucleotides in length.

2. Designing primer pairs for co-amplification of the selected loci.

Primers can be designed using standard methods known in the art.

In some embodiments, designing suitable primer pairs can include determining the efficiency of the PCR reaction on whole blood DNA in simplex (separate) reactions.

- 15 Methods for determining efficiency of PCR are known in the art. For example, efficiency can be determined by (i) selecting a particular concentration of the primers; (ii) conducting real-time PCR reactions on serial dilutions of a template DNA (e.g., whole blood DNA) and determining C_q values for each reaction (each dilution); (iii) generating a standard curve by plotting the C_q values against the log of starting quantity of the template for each dilution; (iv) calculating a slope for the standard curve;
20 and (v) determining reaction efficiency based on the slope. Typically, the PCR efficiency is calculated using the following formula: $\text{Efficiency} = 10^{(-1/\text{slope})} - 1$. Amplification efficiency is also frequently presented as a percentage, that is, the percent of template that was amplified in each cycle; this is given by $\% \text{efficiency} = \text{Efficiency} \times 100$. The factor of amplification is given by $10^{(-1/\text{slope})}$. A %efficiency of 100% (corresponding to an Efficiency of 1 and a slope of -3.32) is considered to reflect perfect
25 efficiency. A %efficiency between 90% and 110% is considered optimal. %Efficiency values which are lower than 90% can be a result of, e.g., poor primer design or non-optimal reaction conditions or non-optimal reagent concentrations. %efficiency values which are higher than 110% (which is observed when the addition of more DNA template does not result in threshold C_q values being reached at earlier cycles) can be a result of, e.g., polymerase inhibition, cross-contamination, the presence of
30 primer dimers. As used herein, the term "equal efficiency" with respect to amplification efficiency/primer efficiency refers to the exact same efficiency (e.g., the same percentage efficiency) and also to differences of up to 5% in efficiency.

It is not always necessary to know the primer efficiency. In some embodiments, designing suitable primer pairs can include determining the amount of PCR products obtained using the primers on whole

blood DNA in simplex (separate) reactions. For example, the level of amplification of the PCR products can be measured using quantitative dyes such as SYBR. Measuring the amount of PCR product obtained with a particular pair of primers bypasses the need to calculate primer efficiency.

3. Checking that the loci pairs are suitable for the assay of the present invention

- 5 In some embodiments, the primers for the genomic loci can be “calibrated” until the same PCR efficiency is achieved. The PCR efficiency can be modified by slightly changing the sequences of the primers, for example by adding/ deleting bases from the 5' or 3' ends of the primers. Alternatively, or additionally, primer concentration can be adjusted in order to achieve equal PCR efficiency.

10 In other embodiments, the primers are tested for their ability to amplify the genomic loci to obtain a comparable level of amplicons, e.g., using a quantitative dye such as SYBR.

To check that the selected loci pairs are able to distinguish between plasma DNA samples and whole blood DNA samples, the first and second loci may be co-amplified from a plasma sample and a whole blood sample to obtain Cq values for each of the loci and for each of the samples. The deltaCQ (ΔCQ) between the first locus and the second locus may be calculated as follows: $\Delta Cq = Cq \text{ (first locus)} - Cq \text{ (second locus)}$. Suitable pairs of loci are those which have a difference between ΔCq in plasma DNA and ΔCq in whole blood DNA of at least 1 cycle.

Method using one locus

20 It is also possible to perform the method using only one locus. In some embodiments, the locus is present at a higher concentration in cfDNA relative to whole blood DNA (a ‘High’ locus). In other embodiments, the locus is present at a lower concentration in cfDNA relative to whole blood DNA (a ‘Low’ locus).

25 When a ‘High’ locus is amplified under certain conditions, the number of amplicons obtained is higher for a plasma sample than for a whole blood sample. When a ‘Low’ locus is amplified under certain conditions, the number of amplicons obtained is lower for a plasma sample than for a whole blood sample.

30 For a locus which is present at a different concentration in cfDNA relative to whole blood DNA, the concentration of that locus in a sample is a reflection of the relative amounts of cfDNA and whole blood DNA contained in the sample. Accordingly, the efficiency of separation can be inferred from the concentration of the locus. When such a locus is amplified, the number of amplicons produced also reflects the relative amounts of plasma DNA and whole blood DNA in the sample, and accordingly, the efficiency of separation of the plasma can be inferred from the amount of amplicons obtained from the locus. In some embodiments, the difference in concentrations of the locus in cfDNA and whole blood DNA is at least 1.5-fold. In some embodiments, the difference in concentrations of the locus in cfDNA and whole blood DNA is at least 2-fold. In some embodiments, the difference in concentrations

of the locus in cfDNA and whole blood DNA is at least 3-fold. In some embodiments, the difference in concentrations of the locus in cfDNA and whole blood DNA is at least 4-fold. In some embodiments, the difference in concentrations of the locus in cfDNA and whole blood DNA is more than 2-fold.

5 In some embodiments, the method according to the invention comprises: (a) generating from a plasma DNA sample, by PCR amplification, a first amplification product from a first genomic locus using a first primer pair, wherein the first locus is present at different concentrations in cfDNA and whole blood DNA; (b) quantitating the amplification product; and (c) determining if the plasma sample is separated based on the quantitated amplification product.

10 In some embodiments, step (b) includes determining a quantification cycle (Cq) for the first amplification product, and the plasma sample is determined to be separated based on a comparison of the Cq value of the first amplification product to a predefined threshold.

In some embodiments, the method includes quantitating the absolute amount of DNA for amplification before step (a).

15 In some embodiments, the first genomic locus is a 'High' locus. In other embodiments, the first genomic locus is a 'Low' locus.

A suitable genomic locus can be selected by selecting a genomic locus which has a different concentration in cfDNA than in whole blood DNA.

20 This may be achieved by selecting a 'peak' or 'trough' region within a whole genome sequencing trace of plasma DNA that plots read count (i.e., sequencing depth) as a function of the genomic locus. It may also be achieved by comparing such whole genome sequencing traces of plasma DNA and whole blood DNA and selecting loci for which the read count differs between plasma DNA and whole blood DNA. The whole genome sequencing read depth may be expressed using HitspanN, where N is a positive even integer (such as Hitspan100).

25 Selecting a suitable genomic locus may also be achieved by using nucleosome positioning data to infer which loci are likely have a different concentration in plasma DNA relative to whole blood DNA. The skilled person will understand that any data which might indicate variations in concentration of loci within plasma DNA, or between plasma DNA and whole blood DNA, can be exploited to select a genomic locus that may be suitable for use in the present invention.

30 To check that the selected locus is able to distinguish between plasma DNA samples and whole blood DNA samples, the locus is amplified from a plasma sample and a whole blood sample to obtain Cq values for each of the samples. A suitable locus is one which has a difference between Cq in plasma DNA and Cq in whole blood DNA of at least 1 cycle.

Exemplary loci, primers and polynucleotide probes

Exemplary loci for determining the efficiency of plasma separation from whole blood include:

- The 'High' locus of SEQ ID NO:1:
5'-GATGACTCCAATAATCAATTACTGAGCTGAATTTGGCTCAGGGAGTATTTGC
CTTTCAATAATGATAGCCTCTTGATAAAGTAAATGCCTTTAGGACCCT-3'
- The 'Low' locus of SEQ ID NO:2:
5'-ATCATCATTTTTAAAACAGCTTCTGCCGTGCTTGGGAAAGGCTGACTTGTGTTA
AGTAATATACGCCCTCTAGTGGATGAAGATGGAGAAGAAAGTATTTGAG-3'
- The 'Reference' locus of SEQ ID NO:3:
5'-AGCAAGGTGAAGACTAAGTCTTCTTGTACAGAATCATCAGGCTAAATTTTT
GGCATTATTTTCAGTCCTTGGAGAC-3'

These correspond to loci in the human genome as specified in Table 1 below.

Table 1: Exemplary loci for use in the present invention (human reference genome build hg38)

Locus type	Locus	Chromosome	Start	End	Amplicon length
'High' locus	SEQ ID NO:1	3	25857401	25857501	101bp
'Low' locus	SEQ ID NO:2	8	81641591	81641692	102bp
'Reference' locus	SEQ ID NO:3	7	121953554	121953630	77bp

Table 2: Exemplary primers and fluorescent probes for use in the present invention

SEQ ID	Description	Sequence	Exemplary fluorophore
SEQ ID NO:4	Forward primer used to amplify SEQ ID NO:1 ('High' locus)	5'-GATGACTCCAATAATCAATTAC-3'	None
SEQ ID NO:5	Reverse primer used to amplify SEQ ID NO:1 ('High' locus)	5'-AGGGTCCTAAAGGCATTACTT-3'	None
SEQ ID NO:10	Fluorescent probe used to detect SEQ ID NO:1 ('High' locus)	5'-AATTTGGCTCAGGGAGTATTTGCC-3'	FAM
SEQ ID NO:6	Forward primer used to amplify SEQ ID NO:2 ('Low' locus)	5'-ATCATCATTTTTAAAACAGCTTCTG-3'	None
SEQ ID NO:7	Reverse primer used to amplify SEQ ID NO:2 ('Low' locus)	5'-CTCAAATACTTTCTTCTCCATCT-3'	None
SEQ ID NO:11	Fluorescent probe used to detect SEQ ID NO:2 ('Low' locus)	5'-AGTAATATACGCCCTCTAGTGGATG-3'	FAM

SEQ ID NO:8	Forward primer used to amplify SEQ ID NO:3 ('Reference' locus)	5' - AGCAAGGTGAAGACTAACTTTTC -3'	None
SEQ ID NO:9	Reverse primer used to amplify SEQ ID NO:3 ('Reference' locus)	5' - GTCTCCAAGGACTGAAATAATGC -3'	None
SEQ ID NO:12	Fluorescent probe used to detect SEQ ID NO:3 ('Reference' locus)	5' - ACAGAATCATCAGGCTAAATTTTGG -3'	JOE

Generation of amplification products

The first and second amplification products of the present invention are generated by amplification using pairs of primers (forward and reverse) designed as known in the art to specifically generate each amplification product.

In some embodiments, the first amplification product comprises a sequence according to SEQ ID NO: 1. In some embodiments, it consists of the sequence according to SEQ ID NO: 1.

In some embodiments, the first amplification product comprises a sequence according to SEQ ID NO: 2. In some embodiments, it consists of the sequence according to SEQ ID NO: 2.

10 In some embodiments, the second amplification product comprises a sequence according to SEQ ID NO: 3. In some embodiments, it consists of the sequence according to SEQ ID NO: 3.

In some embodiments, the first and second amplification products comprise sequences according to SEQ ID NO: 1 and SEQ ID NO: 3, respectively. In other embodiments, they consist of sequences according to SEQ ID NO: 1 and SEQ ID NO: 3, respectively.

15 In some embodiments, the first and second amplification products comprise sequences according to SEQ ID NO: 2 and SEQ ID NO: 3, respectively. In other embodiments, they consist of sequences according to SEQ ID NO: 2 and SEQ ID NO: 3, respectively.

In some embodiments, the primer pair for amplifying the first amplification product is: SEQ ID NO: 4 (forward) and SEQ ID NO: 5 (reverse). In other embodiments, the primer pair for amplifying the first amplification product is: SEQ ID NO: 6 (forward) and SEQ ID NO: 7 (reverse).

In some embodiments, the primer pair for amplifying the second amplification product is: SEQ ID NO: 8 (forward) and SEQ ID NO: 9 (reverse).

Other primer pairs for amplifying other loci with the characteristics as described in the "Methods using two loci" or "Methods using one locus" section, may also be used in the present invention.

25 Methods may include a step of adding PCR reagents, *e.g.* suitable buffer/salt components (if required in addition to buffer/salt remaining from digestion), a DNA polymerase (such as a Taq polymerase), dNTPs, primers and (optionally) probes. The reaction may be a hot-start PCR reaction, in which the

polymerase reaction is initiated at or above the primer annealing temperature. Hot-start PCR is advantageous in preventing extension of primers which anneal non-specifically at ambient temperatures (e.g., 20-55C), at which Taq polymerase retains partial activity. A hot-start PCR reaction may be designed, for instance, by using modifications that block DNA polymerase activity in low
5 temperatures.

Amplification and detection of amplicons may be carried out by conventional PCR using fluorescently-labeled primers followed by capillary electrophoresis of amplification products. In some embodiments, following amplification the amplification products are separated by capillary electrophoresis and fluorescent signals are quantified. An electropherogram plotting the change in fluorescent signals as a
10 function of size (bp) or time from injection may be generated, wherein each peak in the electropherogram corresponds to the amplification product of a single locus. The peak's height (provided for example using "relative fluorescent units", rFU) may represent the intensity of the signal from the amplified locus. Computer software may be used to detect peaks and calculate the fluorescence intensities (peak heights) of a set of loci whose amplification products were run on the
15 capillary electrophoresis machine, and subsequently the ratios between the signal intensities.

A preferred PCR technique is real-time PCR (also known as qPCR), in which simultaneous amplification and detection of the amplification products are performed. Real-time PCR can be used with non-specific detection or sequence-specific detection. Non-specific detection (e.g. using a dsDNA-binding fluorescent dye, such as SYBR Green) can be used within the methods disclosed
20 herein, but is not ideal if it is desired to distinguish between multiple different amplicons in the same reaction. Thus, it is more typical to use sequence-specific detection, and methods and compositions may use a labelled polynucleotide probe (usually with a fluorophore and fluorescence quencher on the same probe, as in the TaqMan system) which is complementary to a specific sequence within nucleic acid amplicon(s) of interest. Polynucleotide probes may vary in length. In some embodiments, the
25 polynucleotide probes may include between 15-30 bases. Polynucleotide probes may be designed to bind to either strand of the template. Additional considerations include the T_m of the polynucleotide probes, which should preferably be compatible with that of the primers.

The methods disclosed herein can involve simultaneous amplification of more than one target sequence (first and second amplification products) in the same reaction mixture, a process known as multiplex
30 amplification or co-amplification. In order to distinguish between multiple target sequences that are amplified in parallel, polynucleotide probes labeled with distinct fluorescent colours may be used. The dye combinations are chosen to be compatible with the real-time PCR thermocycler of choice. In some embodiments, fluorescence may be monitored during each PCR cycle, providing an amplification plot showing the change of fluorescent signals from the probes as a function of cycle number.

Real-time PCR may thus be achieved by using a hydrolysis probe based on combined reporter and quencher molecules. In such assays, oligonucleotide probes have a fluorescent moiety (fluorophore) attached to their 5' end and a quencher attached to the 3' end. During PCR amplification, the polynucleotide probes selectively hybridize to their target sequences on the template, and as the polymerase replicates the template it also cleaves the polynucleotide probes due to the polymerase's 5'-nuclease activity. When the polynucleotide probes are intact, the close proximity between the quencher and the fluorescent moiety normally results in a low level of background fluorescence. When the polynucleotide probes are cleaved, the quencher is decoupled from the fluorescent moiety, resulting in an increase of intensity of fluorescence. The fluorescent signal correlates with the amount of amplification products, *i.e.* the signal increases as the amplification products accumulate.

Suitable fluorophores include, but are not limited to, fluorescein, FAM, lissamine, phycoerythrin, rhodamine, Cy2, Cy3, Cy3.5, Cy5, Cy5.5, Cy7, FluorX, JOE, HEX, NED, VIC and ROX. Suitable fluorophore/quencher pairs are known in the art, including but not limited to: Iowa Black quencher, FAM-TAMRA, FAM-BHQ1, Yakima Yellow-BHQ1, ATTO550-BHQ2 and ROX-BHQ2.

Fluorescence may be monitored during each PCR cycle, providing an amplification plot showing the change of fluorescent signals from the probe(s) as a function of cycle number. In the context of real-time PCR, the following terminology is used:

"Quantification cycle" ("C_q") refers to the cycle number in which fluorescence increases above a threshold, set automatically by software or manually by the user. In some embodiments, the threshold may be constant for each locus of interest and may be set in advance, prior to carrying out the amplification and detection. In other embodiments, the threshold may be defined separately for each locus after the run, based on the maximum fluorescence level detected for this locus during the amplification cycles.

"Threshold" refers to a value of fluorescence used for C_q determination. In some embodiments, the threshold value may be a value above baseline fluorescence, and/or above background noise, and within the exponential growth phase of the amplification plot.

"Baseline" refers to the initial cycles of PCR where there is little to no change in fluorescence.

Computer software is readily available for analysing amplification plots and determining baseline, threshold and C_q.

As used herein, a "primer" defines an oligonucleotide which is capable of annealing to (hybridizing with) a target sequence, thereby creating a double stranded region which can serve as an initiation point for DNA synthesis under suitable conditions. The terminology "primer pair" refers herein to a pair of primers which are selected to be used together in amplifying a selected nucleic acid sequence by PCR. As commonly known in the art, the primers may be designed to bind to a complementary sequence

under selected conditions. Primers may vary in length, depending on the particular assay format and the particular needs. In some embodiments, the primers may be at least 15 nucleotides long, such as between 15-25 nucleotides or 18-25 nucleotides long. The primers may be adapted to be suited to a chosen amplification system. As commonly known in the art, the primers may be designed by taking
5 into consideration melting temperature of the target sequences, which should preferably be comparable to that of the primers.

In cases of simultaneous amplification of more than one target in the same reaction mixture (co-amplification) using different primer pairs, these different primers may be designed such that they can work at the same annealing temperature during amplification. Thus, primers with similar melting
10 temperature (T_m) can be designed *e.g.* within $\pm 3^\circ\text{-}5^\circ\text{C}$ of each other.

Computer software is readily available for routine designing of primers and probes which meet the various requirements of any particular experiment.

Determining efficiency of separation

The efficiency of plasma separation is determined based on the quantitated amplification product(s).
15 The skilled person will understand that any measure of the amount of a sequence-specific amplification product, which reflects the initial amount of copies of the target sequence, may be used. However, the quantitation of amplification products may not involve calculation of any absolute amounts of amplification products/target sequences. Thus, in some embodiments, no reference DNA is employed for quantitating amplification products, as it is unnecessary to calculate actual DNA amounts *per se*.

20 In some embodiments, amplification and detection of amplification products are carried out by real-time PCR where the quantitation of a specific amplification product consists in determining the C_q for this amplification product.

In some embodiments, in case of no amplification or very little amplification the C_q is determined as "infinity". In some embodiments, in such cases the numerical value of the ΔC_q is set to be 14.

25 In some embodiments, plasma may be determined to be efficiently separated following real-time PCR if the difference in the first and second quantitated amplification products is above a predefined threshold. This may comprise determining the C_q for each locus, and calculating the difference between the C_q values (ΔC_q), and determining if the ΔC_q is above a predefined threshold. To evaluate a 'High' locus, calculating the difference in the first and second quantitated amplification products
30 (ΔC_q) is carried out by calculating $C_q(\text{'High' locus}) - C_q(\text{'Reference' locus})$. To evaluate a 'Low' locus, calculating the difference in the first and second quantitated amplification products (ΔC_q) is carried out by calculating $C_q(\text{'Low' locus}) - C_q(\text{'Reference' locus})$.

In other embodiments, the plasma may be determined to be efficiently separated following real-time PCR if the C_q of a 'High' locus is below a certain threshold. In yet other embodiments, the plasma

may be determined to be efficiently separated following real-time PCR if the Cq of a 'Low' locus is above a predefined threshold.

For example, assuming a first ('High' locus) Cq of a first amplification product being 25 and a second Cq of a second ('Reference' locus) amplification product being 30, the Cq('High' locus) - Cq('Reference' locus) is -5. In some embodiments, computer software is used for calculating a difference between the Cq of amplification products.

When a 'High' locus is co-amplified with a 'Reference' locus, a calculated ΔCq indicates that a tested plasma sample is sufficiently separated from whole blood when the calculated ΔCq is below a predefined threshold ΔCq . When a 'Low' locus is co-amplified with a 'Reference' locus, a calculated ΔCq indicates that a tested plasma sample is sufficiently separated from whole blood when the calculated ΔCq is above a predefined threshold ΔCq .

A "threshold Cq" or "threshold ΔCq " refers to a Cq or ΔCq that differentiates between sufficiently separated plasma samples and poorly or not separated plasma samples. The threshold is typically set to reflect contaminating genomic DNA below a certain amount or percentage, which does not interfere with analysis of cell-free DNA. As noted above, different diagnostic applications involving analysis of cell-free DNA may require different levels of purity of the plasma (i.e., may be characterized by different levels of genomic DNA contamination that are tolerated). For example, in some embodiments, a separated plasma sample is a plasma sample containing less than 50% contaminating white blood cell DNA, less than 25% contaminating white blood cell DNA, less than 12.5% contaminating white blood cell DNA, less than 6.25% contaminating white blood cell DNA, less than 3.13% contaminating white blood cell DNA, less than 2.06% contaminating white blood cell DNA, less than 1.03% contaminating white blood cell DNA. Each possibility represents a separate embodiment of the present invention.

A threshold Cq or ΔCq above which a plasma sample is determined to be separated may be set based on the requirements of a particular diagnostic assay.

In some embodiments, the threshold ΔCq is about 1 cycle. In some embodiments, the threshold ΔCq is at least 1 cycle. In additional embodiments, the threshold ΔCq is at least 2 cycles. In yet additional embodiments, the threshold ΔCq is at least 3 cycles. In yet additional embodiments, the threshold ΔCq is at least 4 cycles. The threshold may be less than 20 cycles.

In some embodiments, determining the threshold ΔCq includes:

- (i) spiking a pure plasma sample with different amounts of spiking DNA to obtain a series of spiked plasma samples with different degrees of spiking DNA contamination, wherein the spiking DNA is buffy coat DNA or whole blood DNA;

(ii) generating for each spiked sample, by PCR amplification, a first amplification product from a first genomic locus using a first primer pair, and a second amplification product from a second genomic locus using a second primer pair, wherein the concentration ratio of the first genomic locus to the second genomic locus in cfDNA is different from the concentration ratio of the first genomic locus to the second genomic locus in whole blood DNA;

(iii) calculating a ΔCq for each amplification product and each spiked sample and determining the minimum or maximum threshold according to a chosen degree of spiking DNA contamination.

10 In some embodiments, the lengths (in nucleotides) of the first and second amplification products of step (ii) are within 35% of their mean length; and/or the first and second amplification products of step (ii) are <350 nucleotides in length.

In some embodiments, determining the threshold Cq includes:

(i) spiking a pure plasma sample with different amounts of spiking DNA to obtain a series of spiked plasma samples with different degrees of spiking DNA contamination, wherein the spiking DNA is buffy coat DNA or whole blood DNA;

15 (ii) generating for each spiked sample, by PCR amplification, a first amplification product from a first genomic locus using a first primer pair, wherein the first locus is present at different concentrations in cfDNA and whole blood DNA;

20 (iii) calculating a Cq for the first amplification product and for each spiked sample and determining the minimum or maximum threshold according to a chosen degree of spiking DNA contamination.

In some embodiments, determining the threshold Cq or ΔCq includes using a pure plasma standard as determined by an established method of choice. In some embodiments, determining the threshold Cq or ΔCq includes using a number of biological replicates of plasma sample.

In some embodiments, the methods of the present invention comprise providing a threshold Cq or threshold ΔCq . In some embodiments, the methods of the present invention comprise providing threshold Cq or threshold ΔCq values for multiple different degrees of spiking DNA contamination.

In some embodiments, the efficiency of plasma separation may be determined by comparing a calculated Cq to multiple predefined threshold Cq values corresponding to different degrees of spiking DNA contamination. In other embodiments, the efficiency of plasma separation may be determined by comparing a calculated ΔCq to multiple predefined threshold ΔCq values corresponding to different degrees of spiking DNA contamination.

In some embodiments, the threshold values are statistically significant values. Statistical significance is often determined by comparing two or more populations, and determining a confidence interval (CI) and/or a p value. In some embodiments, the statistically significant values refer to confidence intervals (CI) of about 90%, 95%, 97.5%, 98%, 99%, 99.5%, 99.9% and 99.99%, while preferred p values are less than about 0.1, 0.05, 0.025, 0.02, 0.01, 0.005, 0.001 or less than 0.0001. According to some embodiments, the p value of the threshold value is at most 0.05.

Kits and systems

In some embodiments, a kit is provided for determining the efficiency of plasma separation from whole blood, according to the methods of the present invention.

10 In some embodiments, the kit comprises:

A. primers for amplifying at least one genomic locus, wherein:

- i. the primers comprise a first primer pair for amplifying a first genomic locus, wherein the first locus is present at different concentrations in cfDNA and whole blood DNA; or
- ii. the primers comprise a first primer pair for amplifying a first genomic locus and a
15 second primer pair for amplifying a second genomic locus, wherein the concentration ratio of the first genomic locus to the second genomic locus in cfDNA is different from the concentration ratio of the first genomic locus to the second genomic locus in whole blood DNA and wherein:
 - a. the lengths (in nucleotides) of the first and second amplification products are
20 within 35% of their mean length; and/or
 - b. the first and second amplification products are <350 nucleotides in length;

and B. probes for quantitatively detecting the amplification product(s) produced by the primers of A.

25 In some embodiments, the lengths (in nucleotides) of the first and second amplification products are the same, or within 15% of their mean length.

In some embodiments, the kit further comprises an instruction manual for carrying out the determination of plasma separation efficiency according to the method disclosed herein. In some embodiments, the instruction manual is an electronic instruction manual.

30 In some embodiments, the instruction manual provides a threshold Cq or threshold ΔCq , in comparison to which a plasma sample is determined to be separated. In some embodiments, the instruction manual includes instructions for performing the method steps described herein. In some embodiments, the instruction manual includes instructions directing the correlation between Cq or ΔCq and separation. In some embodiments, the instruction manual includes instructions for carrying out the determination

of plasma separation using a computer software stored on a computer-readable medium, the computer software directs a computer processor to determine plasma separation based on the quantitated amplification product(s).

5 In some embodiments, the kit further comprises a computer readable medium storing a computer software that directs a computer processor to determine plasma separation based on the quantitated amplification product(s).

In additional embodiments, there is provided herein a system for determining the efficiency of plasma separation from whole blood, according to the methods of the present invention.

10 In some embodiments, a system of the present invention comprises primers A as defined above, probes B as defined above, and C. computer software stored on a computer readable medium that directs a computer processor to determine plasma separation based on the quantitated amplification product(s).

15 In some embodiments, the computer software of the present invention directs the computer processor to quantitate the amplification product(s). In some embodiments, the computer software further directs the computer processor to compare the quantitate the amplification product(s) to a predefined threshold, and, based on the comparison, to output whether the plasma sample is separated. In some embodiments, the computer software may be a computer software that directs a computer processor to calculate at least one of C_q and ΔC_q .

20 In some embodiments, the computer software receives as an input parameters or raw data of a real-time PCR run. In some embodiments, the computer software directs a computer processor to analyze the real-time PCR run to quantitate the amplification product(s) (C_q s). In some embodiments, the computer software directs a computer processor to analyze the real-time PCR run to quantitate the amplification product (C_q). In some embodiments, the computer software directs a computer processor to analyze the real-time PCR run to quantitate the amplification products (C_q s) and then calculate the difference in quantitated amplification products (ΔC_q).

25 The computer software includes processor-executable instructions that are stored on a non-transitory computer readable medium. The computer software may also include stored data. The computer readable medium is a tangible computer readable medium, such as a compact disc (CD), magnetic storage, optical storage, random access memory (RAM), read only memory (ROM), or any other tangible medium.

30 In some embodiments, the system comprises a processor, configured to determine plasma separation based on comparison of a quantitated amplification product(s) to a threshold value.

In some embodiments, the system comprises machinery for carrying out the generation of amplification products, such as a real-time PCR machine.

In some embodiments, the kit or system comprises fluorescently-labeled oligonucleotide probes complementary to sub-sequences within the amplification product(s), for detecting the amplification product(s).

- 5 In some embodiments, the kit or system comprises primer pairs designed to selectively amplify a fragment of the genome to generate amplification product(s), as described in sections “Preferred loci, primers and polynucleotide probes”.

In some embodiments, the kit or system may further include at least one additional ingredient needed for amplification and detection of amplification products, such as DNA polymerase and nucleotide mix.

- 10 In some embodiments, the kit or system may further include suitable reaction buffers and a written protocol for performing the assay. The written protocol may comprise instructions for performing any of the steps disclosed herein, including but not limited to, PCR cycling parameters, Cq determination and analysis, and a Cq or ΔCq threshold.

- 15 It is understood that the computer-related methods, steps, processes described herein are implemented using software stored on non-volatile or non-transitory computer readable instructions that when executed configure or direct a computer processor or computer to perform the instructions.

Method using high-throughput sequencing data

It is also possible to perform the assay of the invention using high-throughput sequencing data (*i.e.* next generation sequencing data).

- 20 As described elsewhere herein, the inventors have observed that the fragmentation patterns in cfDNA result in some loci being present at different concentrations than other loci in plasma DNA. In particular, the concentration of loci in cell-free DNA as a function of genomic position appears to take on a periodic pattern with an approximately 200 bp-period (such as in the HitSpan100 plots of Figs 1A and 6A), which matches nucleosomal spacing. In contrast, the same pattern does not appear to be
25 present in whole blood DNA (see e.g. the HitSpan100 plots of Figs 1B and 6B). Thus, without wishing to be bound by theory, it appears that ‘High’ loci correspond to sequences protected by a nucleosome, whereas ‘Low’ loci correspond to those which are not.

- The number of reads of a ‘High’ locus in high-throughput sequencing data is higher for a plasma sample than for a whole blood sample. The number of reads of a ‘Low’ locus in high-throughput
30 sequencing data is lower for a plasma sample than for a whole blood sample. The read count obtained in high-throughput sequencing data reflects the relative amounts of plasma DNA and whole blood DNA in the sample, and accordingly, the efficiency of separation of the plasma can be inferred from the number of reads obtained for the locus.

Method using two sets of loci from high-throughput sequencing data

In some embodiments, the method according to the invention comprises:

- 5 (a) quantifying, from high-throughput sequencing data of a plasma sample, the read counts of a set of first genomic loci and the read counts of a set of second genomic loci, wherein the ratio of the sum of the concentrations of the genomic loci of the first set of genomic loci to the sum of the concentrations of the genomic loci of the second set of genomic loci in cfDNA is different from the ratio of the sum of the concentrations of the genomic loci of the first set of genomic loci to the sum of the concentrations of the genomic loci of the second set of genomic loci in whole blood DNA; and
- 10 (b) determining the efficiency of plasma separation based on the quantitated read counts.

Typically, the bigger the difference in relative read count between the sum of the first genomic loci and the sum of the second genomic loci in plasma DNA and whole blood DNA, the better the sets of genomic loci are suitable for the assay of the present invention. Therefore, a preferred strategy to find two sets of genomic loci that are suitable for the assay of the present invention is to select a set of genomic loci which are “High” loci and a set of genomic loci which are “Low” loci. Suitable loci may be determined from pure plasma samples using the method of determining two sets of loci, using which a plasma sample may be determined to be efficiently separated as provided elsewhere herein.

15

Any sets of genomic loci, wherein the ratio of the sum of the concentrations of the genomic loci of the first set of genomic loci to the sum of the concentrations of the genomic loci of the second set of genomic loci in cfDNA, is different from that in whole blood DNA, may be used for the assay.

20

The loci typically correspond to single positions in the genome.

In some embodiments, step (b) comprises determining the efficiency of plasma separation based on a nucleosomal plasma separation score, wherein the nucleosomal plasma separation score is calculated by dividing the sum of the read counts of set of first genomic loci by the sum of the read counts of set of second genomic loci.

25

In some embodiments, the read count of a locus is the HitSpanN value of the locus. In some embodiments, N is an even number between 2 and 150. In preferred embodiments, N is 100. In other words, in preferred embodiments, the read count of a locus is the HitSpan100 value of the locus.

The set of first genomic loci can have only 1 genomic locus or it can have >1,000,000 genomic loci, or in between. For example, it may have at least 10, at least 100, at least 1,000, at least 10,000 or at least 100,000 genomic loci.

30

The set of second genomic loci can have only 1 genomic locus or it can have >1,000,000 genomic loci, or in between. For example, it may have at least 10, at least 100, at least 1,000, at least 10,000 or at least 100,000 genomic loci.

Typically, the number of loci in the set of first genomic loci and number of loci in the set of second genomic loci is of the same order of magnitude. For instance, they can both have at least 10, at least 100, at least 1,000, at least 10,000, at least 100,000, or at least 1,000,000 genomic loci.

Generally, the higher the number of loci, the bigger the difference between the ratio of (i) the sum of the concentrations of the genomic loci of the first set of genomic loci to (ii) the sum of the concentrations of the genomic loci of the second set of genomic loci, in plasma vs. whole blood samples. As a result, generally, the higher the number of loci, the smaller the average sample sequencing depth required to be able to discriminate between plasma and whole blood samples using high-throughput sequencing. The inventors have identified a set of about 10 million 'high' loci and a set of about 10 million 'low' loci as described in Example 5A, which can be used to reliably discriminate between plasma and whole blood samples.

In some embodiments, the method includes producing high-throughput sequencing data from a plasma sample. In other embodiments, the method does not include producing the high-throughput sequencing data.

The method does not require changing the sequence of the DNA and enables co-analysis of, e.g., methylation, mutation, copy number and nucleosome positioning, based on the same sequencing data. For example, the method is compatible with methylation analysis of high-throughput sequencing data, which comprises analyzing alignments covering a predefined genomic region of at least 50 bps in length, preferably at least 100 bps in length, that contains a restriction locus of interest, and determining a read count of sequence reads covering the predefined genomic region. Such alignments represent DNA molecules of at least 50 bps in length (preferably at least 100 bps in length), in which the analyzed restriction locus, as well as any additional restriction loci within the DNA molecule, were all methylated in the DNA sample and therefore the DNA molecules remained intact following digestion with the enzymes used in the assay.

While the present method can in principle be carried out using any high-throughput sequencing data, it is advantageous to choose sets of first and second genomic loci that are compatible with the preparation of the DNA from which the high-throughput sequencing data are obtained. For instance, when using high-throughput sequencing data obtained from samples subjected to enzymatic digestion, it may be helpful to exclude loci close to enzyme restriction sites from the present method, to avoid the enzymatic digestion step of the sample preparation having a negative impact on the accuracy of the determination of plasma separation efficiency. In some embodiments, the sets of first and/or second genomic loci are located at least 50 bp, at least 100 bp or at least 200 bp away from restriction sites of

enzymes used for digestion during preparation of the DNA sample used for high-throughput sequencing data. Typically, the sets of first and/or second genomic loci are located at least 100 bp away from restriction sites of enzymes used for digestion during preparation of the high-throughput sequencing data.

- 5 Similarly, it may be helpful to exclude loci located in sex chromosomes, to ensure that the output of the method does not have any dependency from the gender of the organism from which the sample originates. For humans, this means only including loci that are located on chromosomes 1-22 in the first and/or second set of loci. In other words, the sets of first and/or second genomic loci are not located on X or Y chromosomes. It may also be helpful to exclude loci located on mitochondrial DNA,
10 for instance because the number of mitochondrial DNA copies may vary between individuals. Accordingly, in some embodiments, the sets of first and/or second genomic loci are not located on mitochondrial DNA.

- It may be helpful to exclude loci that are particularly under- or overrepresented in the genome (e.g., through gene duplication events), since the read count of such loci can vary between samples and thus
15 may affect the reliability of the nucleosomal plasma separation score as a measure of plasma separation efficiency. Accordingly, in some embodiments, the sets of first and/or second genomic loci have a number of sequence reads between 0.5-1.5 of the median number of sequence reads of all the loci in the high-throughput sequencing data.

- The specific requirements according to which loci should be excluded in the present assay can vary
20 according to the sample preparation associated with the high-throughput sequencing data, and the skilled person will understand that the choice of loci may need to be tailored accordingly.

In some embodiments, the high-throughput sequencing is whole genome high-throughput sequencing. In some embodiments, the high-throughput sequencing is target-specific high-throughput sequencing.

Method using one set of loci from high-throughput sequencing data

- 25 The method according to the invention can also be done using one set of genomic loci and using high-throughput sequencing data. Accordingly, in some embodiments, the method according to the invention comprises:

- (a) quantifying, from high-throughput sequencing data of a plasma sample, the read counts of a set of first genomic loci, wherein the sum of the concentrations of the genomic loci of the
30 first set of genomic loci is present at different concentrations in cfDNA and whole blood DNA; and
(b) determining the efficiency of plasma separation based on the sum of the quantitated read counts.

In some embodiments, step (b) comprises normalizing the read counts determined in step (a) against the median read count of a set of normalising loci in the high-throughput sequencing data, to obtain normalized read counts and determining the efficiency of plasma separation based on the sum of the normalized read counts.

5 A preferred strategy to find a sets of genomic loci suitable for the assay of the present invention is to select a set of genomic loci which are “High” loci or a set of genomic loci which are “Low” loci. Suitable loci may be determined from pure plasma samples using the method of determining a set of loci, using which a plasma sample may be determined to be efficiently separated provided elsewhere herein.

10 The loci typically correspond to a single position in the genome.

In some embodiments, the read count of a locus is the HitSpanN value of the locus. In some embodiments, N is an even number between 2 and 150. In preferred embodiments, N is 100. In other words, in preferred embodiments, the read count of a locus is the HitSpan100 value of the locus.

15 The set of first genomic loci can have only 1 genomic locus or it can have >1,000,000 genomic loci or in between. For example, it may have at least 10, at least 100, at least 1,000, at least 10,000 or at least 100,000 genomic loci.

Generally, the higher the number of loci, the bigger the difference between the sum of the concentrations of the genomic loci of the first set of genomic loci in plasma vs whole blood samples. As a result, generally, the higher the number of loci, the smaller the average sample sequencing depth
20 required to be able to discriminate between plasma and whole blood samples using high-throughput sequencing.

In some embodiments, the method includes producing high-throughput sequencing data from a plasma sample. In other embodiments, the method does not include producing the high-throughput sequencing data. In preferred embodiments, the method does not include producing the high-throughput
25 sequencing data.

The method does not require changing the sequence of the DNA and enables co-analysis of, e.g., methylation, mutation, copy number and nucleosome positioning, based on the same sequencing data. For example, the method is compatible with methylation analysis of high-throughput sequencing data, similarly to what is described for the method using two sets of loci.

30 While the present method can in principle be carried out using any high-throughput sequencing data, it is advantageous to choose a set of first genomic loci that is compatible with the preparation of the DNA from which the high-throughput sequencing data are obtained, for the same reasons as described for the method using two sets of loci.

In some embodiments, the set of first genomic loci and/or the set of normalising loci are located at least 50 bp, at least 100 bp or at least 200 bp away from restriction sites of enzymes used for digestion during preparation of the DNA sample used for high-throughput sequencing data. Typically, the set of first genomic loci and/or the set of normalising loci are located at least 100 bp away from restriction sites of enzymes used for digestion during preparation of the high-throughput sequencing data.

Similarly, it may be helpful to exclude loci located in sex chromosomes, to ensure that the output of the method does not have any dependency from the gender of the organism from which the sample originates. For humans, this means only including loci that are located on chromosomes 1-22 in the set of first genomic loci and/or the set of normalising loci. In other words, the set of first genomic loci and/or the set of normalising loci are not located on X or Y chromosomes. It may also be helpful to exclude loci located on mitochondrial DNA. Accordingly, in some embodiments, the set of first genomic loci and/or the set of normalising loci are not located on mitochondrial DNA.

It may be helpful to exclude loci that are particularly under- or overrepresented in the genome (e.g., through gene duplication events), since the read count of such loci can vary between samples and thus may affect the reliability of the nucleosomal plasma separation score as a measure of plasma separation efficiency. Accordingly, in some embodiments, the set of first genomic loci and/or the set of normalising loci have a number of sequence reads between 0.5-1.5 of the median number of sequence reads of all the loci in the high-throughput sequencing data.

The specific requirements according to which loci should be excluded in the present assay can vary according to the sample preparation associated with the high-throughput sequencing data, and the skilled person will understand that the choice of loci may need to be tailored accordingly.

In some embodiments, the high-throughput sequencing is whole genome high-throughput sequencing. In some embodiments, the high-throughput sequencing is target-specific high-throughput sequencing.

High-throughput sequencing data

Library preparation and sequencing

"High throughput sequencing," (also termed "next generation sequencing") includes sequence determination using methods that determine many (typically thousands to billions) of nucleic acid sequences in parallel. High throughput sequencing generally involves three basic steps: library preparation, sequencing and data analysis. Examples of high throughput sequencing techniques include sequencing-by-synthesis and sequencing-by-ligation (employed, for example, by Illumina Inc., Life Technologies Inc., Roche), nanopore sequencing methods and electronic detection-based methods such as Ion Torrent™ technology (Life Technologies Inc.).

Library preparation for the major high-throughput sequencing platforms requires the ligation of specific adapter oligonucleotides to fragments of the DNA to be sequenced. Sample preparation can

include restriction digestion by methylation-sensitive/dependent restriction endonuclease(s), preferably carried out before adapter ligation to avoid possible digestion of the adapters by the enzymes. The digestion of DNA typically does not result in homogeneous, blunt-ended fragments. Thus, end repair is needed to ensure that each DNA molecule is free of overhangs, and contains 5' phosphate and 3' hydroxyl groups. A typical blunting enzyme mix includes a polymerase and a polynucleotide kinase, for example, T4 DNA polymerase and T4 polynucleotide kinase (PNK). T4 DNA polymerase (in the presence of dNTPs) can fill-in 5' overhangs and trim 3' overhangs down to the dsDNA interface to generate the blunt ends. The T4 PNK can then phosphorylate the 5' terminal nucleotide. For Illumina libraries, incorporation of a non-templated deoxyadenosine 5'-monophosphate (dAMP) onto the 3' end of blunted DNA fragments, a process known as dA-tailing, is also required for library preparation. dA-tails prevent concatamer formation during downstream ligation steps, and enable DNA fragments to be ligated to adapter oligonucleotides with complementary dT-overhangs.

Adapter oligonucleotides, also termed "sequencing adapters", can be ligated to the DNA fragments using end-preserving methods such as enzymatic ligation in which a ligase enzyme covalently links a sequencing adapter to a DNA fragment, making a complete library molecule. Sequencing adapters are ligated at the 5' and 3' ends of DNA fragments in the sequencing library. Sequencing adapters typically include platform-specific sequences for fragment recognition by a particular sequencer: for example, sequences that enable library fragments to bind to the flow cells of Illumina platforms. Each sequencing instrument provider typically uses a specific set of sequences for this purpose.

Sequencing adapters may also include sample indices. "Sample indices", also termed "sample barcodes" are sequences that enable multiple samples to be sequenced together (i.e., multiplexed) on the same instrument flow cell or chip. Each sample index, typically 6–10 bases, is specific to a given sample library and is used for de-multiplexing during data analysis to assign individual sequence reads to the correct sample. Sequencing adapters may contain single or dual sample indexes depending on the number of libraries combined and the level of accuracy desired.

Sequencing adapters may include unique molecular identifiers (UMIs). UMIs are a type of molecular barcodes that provide molecular tracking, error correction and increased accuracy during sequencing. UMIs are short sequences, typically 5 to 20 bases in length, used to uniquely tag each molecule in a sample library. Since each nucleic acid in the starting material is tagged with a unique molecular barcode, bioinformatics software can filter out duplicate reads and PCR errors with a high level of accuracy and report unique reads, removing the identified errors before final data analysis.

In some embodiments, both a sample barcode sequence and a UMI are incorporated into a nucleic acid target molecule.

High-throughput sequencing may be performed using various high-throughput sequencing instruments and platforms, including but not limited to: Novaseq™, Nextseq™ and MiSeq™ (Illumina), 454 Sequencing (Roche), Ion Chef™ (ThermoFisher), SOLiD® (ThermoFisher) and Sequel II™ (Pacific Biosciences). The appropriate platform-designed sequencing adapters are used for preparing the sequencing library.

The high-throughput sequencing may be whole genome sequencing, performed on libraries prepared from endonuclease-treated DNA.

Alternatively, region(s) of interest in the endonuclease-treated DNA can be captured using, for example, a solution-phase or solid-phase hybridization-based process, followed by the high-throughput sequencing. Enrichment of regions of interest followed by high-throughput sequencing is referred to herein as “target-specific high-throughput sequencing”. Target-specific high-throughput sequencing includes, for example, CpG island sequencing and exome sequencing. Target-specific high-throughput sequencing also includes sequencing of specific informative genomic regions, for example, regions known to be differentially methylated between cancer and non-cancer tissues. Capture of genomic regions for target-specific sequencing is typically carried out after library preparation. In some embodiments, the methods disclosed herein comprise enriching genomic regions of interest. In order to preserve the ends of the DNA fragments in the DNA sample (e.g., to allow analysis of sequences starting or ending at nucleotides within restriction loci), enrichment according to the present invention is typically not carried out using PCR amplification of the genomic regions of interest.

Analysis of sequence reads

Nucleotide sequences produced by the sequencing process (“sequence reads”, or simply “reads”), may be mapped against a reference genome. A “reference genome” as used herein refers to a previously identified genome sequence, whether partial or complete, assembled as a representative example of a species or subject. A reference genome is typically haploid, and typically does not represent the genome of a single individual of the species but rather is a mosaic of the genomes of several individuals. A reference genome for the methods of the present invention is typically a human reference genome. In some embodiments, the reference genome is the complete human genome, such as the human genome assemblies available at the website of the National Center for Biotechnology Information (NCBI) or at the University of California, Santa Cruz (UCSC) Genome Browser. An example of a suitable reference genome for human studies is the ‘hg18’ genome assembly. As an alternative, the more recent GRCh38 major assembly can be used (going up to patch p13).

Read mapping is the process to align the reads on a reference genome in order to identify the location of the nucleotide sequences within the reference genome. The sequence reads that align are designated as being “mapped”. The alignment process can maximize the possibility for obtaining regions of sequence identity across the various sequences in the alignment, by allowing mismatches, indels and/or

clipping of some short fragments on the two ends of the reads. The number of reads mapped to a certain genomic locus of interest is referred to herein as the “read count”. Computer software may be used to analyze sequence reads, map sequence reads against a reference genome and quantify the number of reads.

- 5 In some embodiments, the “read count” is the ‘HitspanN’, such as ‘HitSpan100’, as defined elsewhere herein.

In other embodiments, the “read count” is the ‘HitspanN meanM’, as defined elsewhere herein.

Method of determining a maximum or a minimum in a genomic region of sequencing data

- 10 The invention also provides methods to analyse the patterns from sequencing data (e.g. high-throughput sequencing data) such as those which originate from fragmentation patterns in cfDNA and result in loci being present at different concentrations than other loci in plasma DNA. These include methods of determining sets of loci using which a plasma sample may be determined to be efficiently separated.

Accordingly, the invention provides a method of providing a parameter, comprising:

- 15 (a) from sequencing data, defining one or more genomic regions;
(b) for each genomic region, determining a local maximum within the genomic region as the locus having the highest read count.

In some embodiments, the read count in step (b) is the HitSpanN value of the locus. In some embodiments, N is an even number between 2 and 150. Typically, N is 100.

- 20 In some embodiments, the read count in step (b) is the HitSpanN meanM value of the locus. In some embodiments, N is an even number between 2 and 150. In some embodiments, M is an even number between 20 and 100. Typically, N is 100 and M is 50.

Advantageously, HitSpanN meanM values effectively correspond to “smoothed out” HitSpanN values, and so permit accurate determination of local minima and local maxima.

- 25 As mentioned elsewhere herein, the concentration of loci in cell-free DNA as a function of genomic position appears to take on a periodic pattern with a 200 bp-period. Accordingly, in some embodiments, step (a) comprises defining a set of consecutive 200-bp genomic regions from sequencing data.

Provided herein is also a method of locating nucleosome sites, comprising the method of providing a parameter, wherein the parameter is a local maximum.

- 30 The invention also provides a method of providing a parameter, comprising:
(a) from sequencing data, defining one or more genomic regions;

- (b) for each genomic region, determining a local minimum within the genomic region as the locus having the lowest read count.

In some embodiments, the read count in step (b) is the HitSpanN value of the locus. In some embodiments, N is an even number between 2 and 150. Typically, N is 100.

- 5 In some embodiments, the read count in step (b) is the HitSpanN meanM value of the locus. In some embodiments, N is an even number between 2 and 150. In some embodiments, M is an even number between 20 and 100. Typically, N is 100 and M is 50.

In some embodiments, step (a) comprises defining a set of consecutive 200-bp genomic regions from sequencing data.

- 10 The invention also provides a method for training a machine learning algorithm, comprising providing a training dataset which includes as a parameter a local maximum or local minimum as described above.

The invention also provides a method of determining two sets of loci, using which a plasma sample may be determined to be efficiently separated, comprising:

- 15 (i) obtaining high-throughput sequencing data of a pure plasma sample;
(ii) determining a set of first loci using the method of providing a parameter wherein the parameter is a local maximum and determining a set of second loci using the method of providing a parameter wherein the parameter is a local minimum.

- 20 Provided herein is also a method of determining a set of loci, using which a plasma sample may be determined to be efficiently separated, comprising:

- (j) obtaining high-throughput sequencing data of a pure plasma sample;
(ii) determining a set of loci using the method of providing a parameter, wherein the parameter is a local maximum or a local minimum.

Determining efficiency of separation

- 25 The efficiency of plasma separation is determined based on the read count. The skilled person will understand that any measure of the read count, which reflects the amount of copies of the loci in the plasma sample, may be used. However, it is unnecessary to calculate actual DNA amounts *per se*.

In some embodiments, the quantitation of a read counts consists in determining the 'HitSpanN', such as 'HitSpan100'.

- 30 In some embodiments, plasma may be determined to be efficiently separated if the nucleosomal plasma separation score is above or below a certain threshold, and the nucleosomal plasma separation score is calculated by dividing the sum of the read counts of the first genomic loci by the sum of the read counts of the set of second genomic loci.

For example, the plasma may be determined to be efficiently separated if the nucleosomal plasma separation score of a set of 'High' loci is above a certain threshold. In other embodiments, the plasma may be determined to be efficiently separated if the nucleosomal plasma separation score of 'Low' loci is below a predefined threshold.

- 5 By way of further example, a plasma may be determined to be efficiently separated based on a set of 'High' loci and a set of 'Low' loci. If the nucleosomal plasma separation score is calculated by dividing the sum of the read counts of the set of 'High' genomic loci by the sum of the read counts of the set of 'Low' genomic loci, the tested plasma sample is deemed sufficiently separated from whole blood when the calculated nucleosomal plasma separation score is above a predefined threshold value. If the
- 10 nucleosomal plasma separation score is calculated by dividing the sum of the read counts of the set of 'Low' genomic loci by the sum of the read counts of the set of 'High' genomic loci, the tested plasma sample is deemed sufficiently separated from whole blood when the calculated nucleosomal plasma separation score is below a predefined threshold value.

- A "threshold value" refers to a value that differentiates between sufficiently separated plasma samples and poorly or not separated plasma samples. The threshold is typically set to reflect contaminating genomic DNA below a certain amount or percentage, which does not interfere with analysis of cell-free DNA. As noted above, different diagnostic applications involving analysis of cell-free DNA may require different levels of purity of the plasma (i.e., may be characterized by different levels of genomic DNA contamination that are tolerated). For example, in some embodiments, a separated plasma sample
- 15 is a plasma sample containing less than 50% contaminating white blood cell DNA, less than 25% contaminating white blood cell DNA, less than 12.5% contaminating white blood cell DNA, less than 6.25% contaminating white blood cell DNA, less than 3.13% contaminating white blood cell DNA, less than 2.06% contaminating white blood cell DNA, less than 1.03% contaminating white blood cell DNA. Each possibility represents a separate embodiment of the present invention.

- 20 A threshold value above which a plasma sample is determined to be separated may be set based on the requirements of a particular diagnostic assay.

In some embodiments, determining the threshold value includes:

- (i) spiking a pure plasma sample with different amounts of spiking DNA to obtain a series of spiked plasma samples with different degrees of spiking DNA contamination, wherein
- 30 the spiking DNA is buffy coat DNA or whole blood DNA;
- (ii) generating for the pure plasma sample and each spiked sample, high-throughput sequencing data;
- (iii) for each spiked sample, determining the read counts of a set of first genomic loci and the read counts of a set of second genomic loci, wherein the ratio of the sum of the
- 35 concentrations of the genomic loci of the first set of genomic loci to the sum of the

5 concentrations of the genomic loci of the second set of genomic loci in cfDNA is different from the ratio of the sum of the concentrations of the genomic loci of the first set of genomic loci to the sum of the concentrations of the genomic loci of the second set of genomic loci in whole blood DNA, and determining the minimum or maximum threshold value according to a chosen degree of spiking DNA contamination.

In some embodiments, determining the threshold value includes:

- (i) spiking a pure plasma sample with different amounts of spiking DNA to obtain a series of spiked plasma samples with different degrees of spiking DNA contamination, wherein the spiking DNA is buffy coat DNA or whole blood DNA;
- 10 (ii) generating for the pure plasma sample and each spiked sample, high-throughput sequencing data;
- (iii) for each spiked sample, determining the read counts of a set of first genomic loci, wherein the sum of the concentrations of the genomic loci of the first set of genomic loci is present at different concentrations in cfDNA and whole blood DNA, and determining
15 the minimum or maximum threshold according to a chosen degree of spiking DNA contamination.

In some embodiments, determining the threshold value includes using a pure plasma standard as determined by an established method of choice. In some embodiments, determining the threshold value includes using a number of biological replicates of plasma sample.

20 In some embodiments, the methods of the present invention comprise providing a threshold value. In some embodiments, the methods of the present invention comprise providing threshold values for multiple different degrees of spiking DNA contamination.

In some embodiments, the efficiency of plasma separation may be determined by comparing a calculated nucleosomal plasma separation score to multiple predefined threshold values corresponding
25 to different degrees of spiking DNA contamination.

In some embodiments, the threshold values are statistically significant values. Statistical significance is often determined by comparing two or more populations, and determining a confidence interval (CI) and/or a p value. In some embodiments, the statistically significant values refer to confidence intervals (CI) of about 90%, 95%, 97.5%, 98%, 99%, 99.5%, 99.9% and 99.99%, while preferred p values are
30 less than about 0.1, 0.05, 0.025, 0.02, 0.01, 0.005, 0.001 or less than 0.0001. According to some embodiments, the p value of the threshold value is at most 0.05.

Plasma sample processing

As used herein, the terms "whole blood" and "blood" refer to a blood sample that has not been fractionated and contains cellular components (red blood cells, white blood cells, and platelets) as well as fluid components.

- 5 The term "plasma" refers to the liquid remaining after a whole blood sample is subjected to a separation process to remove the blood cells.

In typical embodiments, plasma samples to be analyzed using the methods of the invention are derived from human subjects. According to some embodiments, the plasma samples originate from subjects with a malignant disease such as a particular type of cancer, or subjects suspected of (or potentially) having a malignant disease such as one or more types of cancer. According to additional embodiments, the plasma samples originate from healthy subjects. The term "healthy subjects" as used herein refers to subjects not diagnosed with a malignant disease such as a certain type of cancer and/or subjects not suspected of having cancer.

The plasma samples may be samples separated from whole blood using any method of separation. Methods known in the art include, for example centrifugation or sedimentation. An exemplary procedure is described in the Examples section below. The plasma sample may be a freshly isolated sample or a sample that was stored for a certain period of time before the analysis.

General

The practice of the present invention will employ, unless otherwise indicated, conventional methods of chemistry, biochemistry, and molecular biology, within the skill of the art. Such techniques are explained fully in the literature. See, *e.g.*, *Methods In Enzymology* (Academic Press, Inc.), Green & Sambrook (2012) *Molecular Cloning: A Laboratory Manual*, 4th edition (Cold Spring Harbor Press), Ausubel *et al.* (eds) *Short protocols in molecular biology*, 5th edition (Current Protocols), *Molecular Biology Techniques: An Intensive Laboratory Course*, (Ream & Field, eds., 1998, Academic Press), Wilson and Walker's *Principles and Techniques of Biochemistry and Molecular Biology* (Hodmann & Clokie, 2018), *Basic Molecular Biology & Techniques - Recent Advances: Molecular Biology & Its Technique* (Singh *et al.*, 2021), *etc.*

The term "comprising" encompasses "including" as well as "consisting" *e.g.* a composition "comprising" X may consist exclusively of X or may include something additional *e.g.* X + Y.

- 30 The term "about" in relation to a numerical value x is optional and means, for example, $x \pm 10\%$.

The word "substantially" does not exclude "completely" *e.g.* a composition which is "substantially free" from Y may be completely free from Y. Where necessary, the word "substantially" may be omitted from the definition of the invention.

The term “between” with reference to two values includes those two values *e.g.* the range “between” 10 nt and 20 nt encompasses *inter alia* 10, 15, and 20 nt.

Unless specifically stated, a method comprising a step of mixing two or more components does not require any specific order of mixing. Thus components can be mixed in any order. Where there are three components then two components can be combined with each other, and then the combination may be combined with the third component, etc.

The various steps of methods may be carried out at the same or different times, in the same or different geographical locations, *e.g.* countries, and by the same or different people or entities.

BRIEF DESCRIPTION OF THE DRAWINGS

- Figure 1 shows the number of normalised, Hitspan100 for positions in the human genome, and shows an example of specific ‘High’ and ‘Low’ loci.** Blood samples from human subjects were separated into 2 parts; one part was kept unprocessed (whole blood) and the other part was processed to obtain plasma (see Methods of Example 1). **(A)** shows the number of normalised, overlapping read hits obtained for the plasma sample **(B)** shows the number of normalised, overlapping read hits obtained for the corresponding whole blood sample. Normalisation was achieved by dividing the Hitspan100 value by the median Hitspan100 value of a subset of loci in the sample (*i.e.* loci on chromosomes 1 - 22; >100 bp from any recognition site loci which have a number of sequence reads between 0.5-1.5 of the median number of sequence reads in NPPU and NWBPU samples. NPPU- normal plasma pool, undigested; NWBPU- normal whole blood pool, undigested).
- Figure 2 shows the normalised HitSpan100 values obtained for different ‘High’ and ‘Low’ loci.** **(A)** shows HitSpan100 for a series of loci amplified from the plasma DNA and whole blood DNA samples. “Low1”, “Low2” and “Low3” are three different loci which are depleted in cfDNA relative to whole blood DNA (values from a few dozen biological replicates), and “High1”, “High2” and “High3” are the values obtained for three different loci which are enriched in cfDNA relative to whole blood DNA (values from a few dozen biological replicates). **(B)** shows a table containing the values plotted in A.

- Figure 3 shows exemplary quantitative PCR plots of selected loci in the human genome, quantitated using fluorescently-labelled complementary oligonucleotides.** SEQ ID NOs: 10 and 11 were labelled with FAM and used to quantitate ‘High’ and ‘Low’ loci, respectively. SEQ ID NO: 12 was labelled with JOE and was used to detect the ‘Reference’ locus. **(A)** shows a quantitative PCR plot of the co-amplification of the ‘High’ locus of SEQ ID NO:1 and the ‘Reference’ locus of SEQ ID NO: 3 from a human plasma sample. **(B)** shows a quantitative PCR plot of the co-amplification of the ‘High’ locus of SEQ ID NO:1 and the ‘Reference’ locus of SEQ ID NO: 3 from a human whole blood sample. **(C)** shows a quantitative PCR plot of the co-amplification of the ‘Low’ locus of SEQ ID NO:2 and the ‘Reference’ locus of SEQ ID NO: 3 from a human plasma sample. **(D)** shows a quantitative

PCR plot of the co-amplification of the 'Low' locus of SEQ ID NO:2 and the 'Reference' locus of SEQ ID NO: 3 from a human whole blood sample.

Figure 4 shows ΔCQ values obtained for different mixtures of plasma and whole blood DNA.

Contrived DNA samples containing plasma DNA with increasing amounts of whole blood DNA were tested to obtain a measure of the assay sensitivity. (A) shows ΔCQ values obtained from the co-amplification of the 'High' locus of SEQ ID NO:1 and 'Reference' locus of SEQ ID NO:3, measured from mixtures of 100% plasma DNA, 75% plasma DNA and 25% whole blood DNA, 50% plasma DNA and 50% whole blood DNA, 25% plasma DNA and 75% whole blood DNA, and 100% whole blood DNA. (B). Same as A, but from the co-amplification of the 'Low' locus of SEQ ID NO:2 and 'Reference' locus of SEQ ID NO:3.

Figure 5 shows the ΔCQ values obtained for the loci of interest for both human plasma and human whole blood samples. This figure demonstrates the full segregation between ΔCQ values obtained for plasma and whole blood samples, for both 'High' and 'Low' loci. (A) shows the ΔCQ values obtained from the co-amplification of the 'High' locus of SEQ ID NO:1 and 'Reference' locus of SEQ ID NO:3, for 32 plasma samples obtained from different human subjects and 35 whole blood samples obtained from different human subjects. (B) Same as A, but obtained from the co-amplification of the 'Low' locus of SEQ ID NO:2 and 'Reference' locus of SEQ ID NO:3.

Figure 6 shows HitSpan100 values obtained for of a virtual pool of plasma samples and a virtual pool of whole blood samples, and the determination of the set of 'low' loci and the set of 'high' loci used in the nucleosomal plasma separation score. Chromosome 15, positions 61165844-61168195 are shown in this figure as an exemplary genomic region. **Virtual pool of plasma samples:** (A) 'HitSpan100' values obtained for the exemplary genomic region. (B) 'HitSpan100 mean50' values for each site of the exemplary genomic region (obtained by averaging the HitSpan100 values over a window spanning 25 nucleotides upstream to 25 downstream of the site). (C) For each successive 200-bp window, the genomic locus with the maximum HitSpan100 mean50 value is plotted ('HitSpan100 mean50 max200' value). (D) For each successive 200-bp window, the genomic locus with the minimum HitSpan100 mean50 value is plotted ('HitSpan100 mean50 min200' value). **Virtual pool of whole blood samples:** (E) 'HitSpan100' values obtained for the exemplary genomic region. (F) 'HitSpan100 mean50' values obtained for each site of the exemplary genomic region (by averaging the HitSpan100 values over a window spanning 25 nucleotides upstream to 25 downstream of each site). (G) For each successive 200-bp window, the genomic locus with the maximum HitSpan100 mean50 value is plotted ('HitSpan100 mean50 max200' value). (H) For each successive 200-bp window, the genomic locus with the minimum HitSpan100 mean50 value is plotted ('HitSpan100 mean50 min200' value).

Figure 6 shows HitSpan100 values obtained for of a virtual pool of plasma samples and a virtual pool of whole blood samples, and the determination of the set of 'low' loci and the set of 'high' loci used in the nucleosomal plasma separation score. Chromosome 15, positions 61165844-61168195 are shown in this figure as an exemplary genomic region. **Virtual pool of plasma samples:** (A) 'HitSpan100' values obtained for the exemplary genomic region. (B) 'HitSpan100 mean50' values for each site of the exemplary genomic region (obtained by averaging the HitSpan100 values over a window spanning 25 nucleotides upstream to 25 downstream of the site). (C) For each successive 200-bp window, the genomic locus with the maximum HitSpan100 mean50 value is plotted ('HitSpan100 mean50 max200' value). (D) For each successive 200-bp window, the genomic locus with the minimum HitSpan100 mean50 value is plotted ('HitSpan100 mean50 min200' value). **Virtual pool of whole blood samples:** (E) 'HitSpan100' values obtained for the exemplary genomic region. (F) 'HitSpan100 mean50' values obtained for each site of the exemplary genomic region (by averaging the HitSpan100 values over a window spanning 25 nucleotides upstream to 25 downstream of each site). (G) For each successive 200-bp window, the genomic locus with the maximum HitSpan100 mean50 value is plotted ('HitSpan100 mean50 max200' value). (H) For each successive 200-bp window, the genomic locus with the minimum HitSpan100 mean50 value is plotted ('HitSpan100 mean50 min200' value).

Figure 7 shows the ability of the nucleosomal plasma separation score to distinguish between whole blood and plasma samples. (A) Box plot of nucleosomal plasma separation score in whole blood and plasma samples from different cohorts 1-3. Scores of whole blood samples are approximately 1, whereas the scores of plasma samples from the 3 cohorts are larger than 1. There is a 100% separation between the whole blood scores and the plasma scores. The p-value was calculated based on a T-test. **(B)** shows a table containing the values plotted in A. Mean, standard deviation and range of nucleosomal plasma separation score are shown for each cohort. “All plasma” includes the data for plasma cohorts 1-3.

EXAMPLES

10 *Example 1: The patterns of amplification using whole genome sequencing are different for plasma and whole blood samples*

In this example, we tested whether the coverage depth in whole-genome sequencing differs between plasma and whole blood DNA samples.

Results

15 Blood samples were collected from human subjects. Samples were separated into 2 parts; one part was kept unprocessed (whole blood) and the other part was processed to obtain plasma (see Methods). DNA was extracted from each of the 2 samples (plasma and whole blood; see Methods).

Both DNA samples were subjected to whole genome sequencing. The HitSpan100 is shown for the plasma DNA and whole blood DNA samples.

20 As shown in Figure 1, the Hitspan100 differs substantially between plasma DNA and whole blood DNA across the same genomic loci. While some loci appear to be sequenced at about 3x higher depth than other loci for plasma DNA, the read depth of whole blood DNA appears to be a lot more uniform.

Conclusion

25 Plasma DNA sequencing depth is not uniform across the genome: a ‘High’ locus is enriched in a plasma sample relative to a whole blood sample, and a ‘Low’ locus is depleted in a plasma sample relative to a whole blood sample. The Hitspan100 measurable informs on the read depth of entire loci which are 100bp long, which is advantageous for identifying loci which can be exploited in the present assay. Such loci were evaluated in Examples 3 and 4 in order to develop assays which determine the efficiency of plasma separation.

30 Methods

Separation of plasma from blood

Blood tubes containing anticoagulant were centrifuged at 1,500 g for 10 minutes. During this first centrifugation step, the blood components are separated. The plasma (without the buffy coat layer) was

then transferred into a new tube and centrifuged again (1,500 g for 10 minutes). The pure plasma was transferred into a new tube and used for downstream processes.

DNA Extraction from plasma

DNA was extracted from plasma samples using a QIAamp Circulating Nucleic Acid Kit (QIAGEN).

5 *DNA Extraction from whole blood*

DNA was extracted from whole blood samples using a QIAamp DNA mini kit (QIAGEN).

Example 2: Fragmentomics-based plasma separation signal in cell-free DNA and whole blood DNA using co-amplification of two loci

Using the methods described for Example 1, we tested the coverage depth in whole-genome sequencing for 'Low' and 'High' loci.
10

Results

Blood samples were collected from human subjects. Samples were each separated into 2 parts; one part was kept unprocessed (whole blood) and the other part was processed to obtain plasma (see Methods). DNA was extracted from each of the 2 samples (plasma and whole blood; see Methods).

15 All DNA samples were subjected to whole genome sequencing.

In Figure 2A, the HitSpan100 is plotted for a series of different loci amplified from the plasma DNA and whole blood DNA samples. "Low1", "Low2" and "Low3" are three different loci which are depleted in cfDNA relative to whole blood DNA (values from a few dozen biological replicates), and "High1", "High2" and "High3" are three different loci which are enriched in cfDNA relative to whole blood DNA (values from a few dozen biological replicates).
20

The HitSpan100 values were insignificantly different for samples obtained from human subjects that had been diagnosed with lung cancer.

Conclusions

The inventors were able to identify various loci that differ in HitSpan100 values between plasma DNA and whole blood DNA samples and which can be exploited in the present invention.
25

Example 3: Fragmentomics-based plasma separation signal in cell-free DNA and whole blood DNA

Based on the findings from Example 1, an assay is designed to evaluate the efficiency of plasma separation.

Whole blood and plasma samples are obtained as described in Example 1. The plasma DNA is extracted from each sample as described in Example 1.
30

Example 3A: Evaluation of a 'High' locus for determining plasma separation

For each of the DNA samples, the 'High' locus of SEQ ID NO:1 is amplified using real-time PCR and detected using FAM fluorescence (see Methods). A Cq value is obtained for the plasma sample which is a lower value than the Cq value obtained for the whole blood sample. These Cq values are taken forward as reference values against which the efficiency of plasma separation of an uncharacterised plasma sample can be determined.

As the plasma sample serves as a reference against which an uncharacterised sample can be assessed, it may be useful to assess its purity. The purity of the plasma sample can be ascertained using another method (e.g., by using a fragment analyser) (to a level of purity as required by the downstream applications, which may require different levels of purity of the plasma). Alternatively, the present example can be carried out with a number of biological plasma sample replicates, from which an mean reference Cq may be determined. The existing method (which has drawbacks relative to the present method as explained in the Background section) is only used for the sole purpose of setting up the present fragmentomics-based assay.

15 Determining plasma separation for an uncharacterised plasma sample

An uncharacterised plasma sample is provided, for which the degree of whole blood DNA contamination is unknown. The plasma DNA is extracted as described in Example 1. The 'High' locus of SEQ ID NO:1 is amplified using real-time PCR and detected using FAM fluorescence. A Cq value is obtained for the uncharacterised sample. The degree of contamination of whole blood DNA is inferred by rule of proportionality, by comparing its Cq value against the reference Cq values obtained for the pure plasma and whole blood samples.

Example 3B: Evaluation of a 'Low' locus for determining plasma separation

For each of the DNA samples, the 'Low' locus of SEQ ID NO:2 is amplified using real-time PCR and detected using FAM fluorescence (see Methods). A Cq value is obtained for the plasma sample which is a lower value than the Cq value obtained for the whole blood sample. These Cq values are taken forward as reference values against which the efficiency of plasma separation of an uncharacterised plasma sample can be determined.

As the plasma sample serves as a reference against which an uncharacterised sample can be assessed, it may be useful to assess its purity, as in Example 3A.

30 Determining plasma separation for an uncharacterised plasma sample

An uncharacterised plasma sample is provided, for which the degree of whole blood DNA contamination is unknown. The plasma DNA is extracted as described in Example 1. The 'Low' locus of SEQ ID NO:2 is amplified using real-time PCR and detected using FAM fluorescence. A Cq value is obtained for the uncharacterised sample. The degree of contamination of whole blood DNA in the

uncharacterised sample is inferred by rule of proportionality, by comparing its Cq value against the reference Cq values obtained for the pure plasma and whole blood samples.

Example 3C: Evaluation of a ‘High’ locus for determining plasma separation using a standard curve

5 A series of mixtures is generated from the pure plasma DNA and whole blood DNA samples. The final samples are 100% plasma DNA, 75% plasma DNA and 25% whole blood DNA, 50% plasma DNA and 50% whole blood DNA, 25% plasma DNA and 75% whole blood DNA, and 100% whole blood DNA. For each of the DNA samples, the ‘High’ locus of SEQ ID NO:1 is amplified using real-time PCR and detected using FAM fluorescence (see Methods), and Cq values are obtained. A standard
10 curve of Cq values against degree of whole blood contamination is constructed.

As the plasma sample serves as a reference against which an uncharacterised sample can be assessed, it may be useful to assess its purity, as in Example 3A.

Determining plasma separation for an uncharacterised plasma sample

15 An uncharacterised plasma sample is provided, for which the degree of whole blood DNA contamination is unknown. The plasma DNA is extracted as described in Example 1. The ‘High’ locus of SEQ ID NO:1 is amplified using real-time PCR and detected using FAM fluorescence. A Cq value is obtained for the uncharacterised sample. The degree of contamination of whole blood DNA in the uncharacterised sample is inferred from the standard curve constructed using samples of known purity.

The skilled person would understand that Example 3C can equally be carried out using a ‘Low’ locus,
20 for example the locus of SEQ ID: NO 2.

Conclusions

This example establishes a rapid assay for determining the purity of plasma samples. This assay can either be performed using a ‘High’ locus (which is enriched in plasma samples relative to whole blood samples) or a ‘Low’ locus (which is depleted in plasma samples relative to whole blood samples). By
25 using a pure plasma sample (e.g., as determined from an existing method) the assay is calibrated using a reliable sample. Advantageously, the existing method is only required for setting up the assay. The purity of any subsequent plasma sample can then be determined using the present invention: The ‘High’ or ‘Low’ locus of that subsequent sample is amplified, and then the resulting Cq value is compared against those of the pure plasma and whole blood cell samples (either by rule of
30 proportionality as described in Example 3A or 3B or using a full standard curve as described in Example 3C). Compared to existing methods, the present method is not sample intensive, it is rapid and it is reliable.

Methods

Real time PCR amplification and detection

Loci are amplified using real-time PCR.

The amplification reaction (total volume 25 μ L) contains 3 ng of the extracted DNA, 0.2 μ M primers, dNTPs and a reaction buffer.

To enable quantitative detection of amplification products during amplification, for each amplicon of interest, a fluorescently-labelled complementary oligonucleotide probes is added to the reaction. Table 1 contains the sequences of the fluorescently-labelled oligonucleotide probes. SEQ ID NOs: 10 and 11 are labelled with FAM.

Real-time PCR reactions were carried out in an ABI 7500 FastDx instrument. PCR conditions are as follows:

Step 1: 95°C, 10 min.

Step 2: 45 cycles of: (i) 95°C, 15 sec; (ii) 60°C, 1 min.

Example 4: Fragmentomics-based plasma separation signal in cell-free DNA and whole blood DNA using co-amplification of two loci

In this example, a method was developed using co-amplification of two loci in order to distinguish between cell-free DNA obtained from human plasma, and whole blood DNA obtained from human whole blood samples.

Example 4A: Evaluation of pairs of loci for determining plasma separation

For each of the DNA samples, the 'High' locus of SEQ ID NO:1 and 'Reference' locus of SEQ ID NO:3 were co-amplified using real-time PCR, and detected using FAM and JOE fluorescence, respectively (see Methods). Cq values were obtained for the plasma sample and whole blood samples for each of the loci. The deltaCQ (Δ CQ) between the 'High' locus and the 'Reference' locus was calculated for each DNA sample as follows: Δ CQ = CQ ('High' locus) - CQ ('Reference' locus). Figure 3A shows the real-time PCR traces for the 'High' locus in the plasma sample (Δ CQ = -2.23) and Figure 3B shows the real-time PCR traces for the 'High' locus in the whole blood sample (Δ CQ = -0.73).

Next, an alternative pair of loci was evaluated. For each of the DNA samples, the 'Low' locus of SEQ ID NO:2 and 'Reference' locus of SEQ ID NO:3 were co-amplified using real-time PCR, and detected using FAM and JOE fluorescence, respectively (see Methods). Cq values were obtained for the plasma sample and whole blood samples for each of the loci. The deltaCQ (Δ CQ) between the 'Low' locus and the 'Reference' locus was calculated for each DNA sample as follows: Δ CQ = CQ ('Low' locus) - CQ ('Reference' locus). Figure 3C shows the real-time PCR traces for the 'Low' locus in the plasma sample (Δ CQ = 1.8) and Figure 3D shows the real-time PCR traces for the 'Low' locus in the whole blood sample (Δ CQ = -0.51).

Example 4B: Standard curve using pairs of loci for determining plasma separation

A series of mixtures was generated from the pure plasma DNA and whole blood DNA samples. These samples are 100% plasma DNA, 75% plasma and 25% whole blood DNA, 50% plasma and 50% whole blood DNA, 25% plasma and 75% whole blood DNA, and 100% whole blood DNA.

- 5 Again, the 'High' locus of SEQ ID NO:1 and 'Reference' locus of SEQ ID NO:3 were co-amplified using real-time PCR, and detected using FAM and JOE fluorescence, respectively (see Methods). Δ CQ values were obtained for each sample and are plotted in Figure 4A. A linear fit of these data gave rise to a R^2 value of 0.9815.

- 10 Analogously, the 'Low' locus of SEQ ID NO:2 and 'Reference' locus of SEQ ID NO:3 were co-amplified using real-time PCR, and detected using FAM and JOE fluorescence, respectively (see Methods). Δ CQ values were obtained for each sample and are plotted in Figure 4B. A linear fit of these data gave rise to a R^2 value of 0.9731.

Example 4C: Reproducibility of Δ CQ values

- 15 Next, the reproducibility of the Δ CQ values obtained for plasma samples and blood samples was verified. 32 human plasma samples and 35 human whole blood samples were prepared according to the same method, and the DNA was extracted from each sample.

- The Δ CQ values obtained by co-amplification of the 'High' locus of SEQ ID NO:1 and 'Reference' locus of SEQ ID NO:3 are plotted for each plasma sample and each whole blood sample in Figure 5A. The Δ CQ values obtained by co-amplification of the 'Low' locus of SEQ ID NO:2 and 'Reference' locus of SEQ ID NO:3 are plotted for each plasma sample and each whole blood sample in Figure 5B.

Conclusions

- 25 This example demonstrates that 2 loci can be used together in order to discriminate between a plasma and whole blood DNA sample. Relative to Example 3, the use of two loci removes the need for measuring DNA concentration before amplifying the loci. The second locus in effect serves as an internal control for DNA concentration.

It is also possible to construct a full standard curve using mixtures of plasma and whole blood DNA, as demonstrated in Example 4B. Lastly, Example 4C demonstrates that the Δ CQ values obtained by co-amplification of these loci are highly reproducible, attesting to the robustness of the method.

30 **Methods**

Real time PCR amplification and detection

Loci were co-amplified using real-time PCR.

The amplification reaction (total volume 25 μ L) contained 2 μ L of the extracted DNA (DNA concentration was not measured before amplification), 0.05-0.5 μ M primers, dNTPs and a reaction buffer.

To enable quantitative detection of amplification products during amplification, for each amplicon of interest, a fluorescently-labeled complementary oligonucleotide probes was added to the reaction. Table 2 contains the sequences of the fluorescently-labeled oligonucleotide probes. SEQ ID NOs: 10 and 11 were labelled with FAM and SEQ ID NO: 12 was labelled with JOE.

Real-time PCR reactions were carried out in an ABI 7500 FastDx instrument. PCR conditions were as follows:

Step 1: 95°C, 10 min.

Step 2: 45 cycles of: (i) 95°C, 15 sec; (ii) 60°C, 1 min.

Example 5: Development of a nucleosomal plasma separation score for determining the purity of plasma samples using next generation sequencing (NGS) data

In this example, a method was developed to distinguish between cell-free DNA obtained from human plasma, and whole blood DNA obtained from human whole blood samples based on features extracted from NGS data.

Example 5A: Development of the assay

Virtual pools of plasma DNA and whole blood DNA were created from individual sequencing libraries. The libraries were typically prepared using standard kits such as NEBNext® Ultra™ II DNA Library Prep Kit for Illumina. From these data, HitSpan100 values (i.e. the number of sequence reads which span the site (i.e. single nucleotide position) with at least 50 nucleotides both upstream and downstream) were obtained for virtual pooled plasma DNA as shown in Figure 6A, and for virtual pooled whole blood DNA as shown in Figure 6E.

Then, for each of the virtual pooled plasma and virtual pooled whole blood DNA samples, “HitSpan100 mean50” values were determined for each site (i.e. single nucleotide position) by averaging the HitSpan100 values over a window spanning 25 nucleotides upstream to 25 downstream of said site (Figures 6B and 6F). The “HitSpan100 mean50” values effectively correspond to “smoothed out” HitSpan100 values, and so permit accurate determination of local minima and local maxima. An initial set of ‘high’ genomic loci was identified as those genomic loci with the maximum height of “HitSpan100 mean50” value over successive 200 bp windows (Figures 6C and 6G), and an initial set of ‘low’ genomic loci were identified as those genomic loci with the minimum height of “HitSpan100 mean50” values over successive 200 bp windows (Figures 6D and 6H). From the initial sets of ‘high’ and ‘low’ genomic loci, final sets of ‘high’ and ‘low’ genomic loci were determined, using the following criteria:

- Only those loci from chromosomes 1-22 were included (i.e., loci on chromosome X, chromosome Y and mitochondrial DNA were excluded), to ensure that the score is reproducible across different human samples (e.g. from males and females).
- Only loci >100 bp from restriction sites of enzymes used to digest the DNA samples were included, to avoid the digestion step having any impact on the nucleosomal plasma separation score.
- Only loci which gave rise to read counts > 0.5 and <1.5 of the median read count in NPPU and NWBPU samples were included (NPPU- normal plasma pool, undigested; NWBPU- normal whole blood pool, undigested). This was done to exclude over- and under-represented genomic loci in the score, because the read count of such loci can vary between samples and thus may affect the reliability of the nucleosomal plasma separation score as a measure of plasma separation efficiency.

Once determined, these final sets of 'high' and 'low' genomic loci were used throughout the present example. The final set of 'low' genomic loci consisted of 9,832,078 loci and the final set of 'high' genomic loci consisted of 9,761,845 loci.

A nucleosomal plasma separation score was then calculated as the sum of the "HitSpan100" values of the final set of 'high' genomic loci divided by the sum of "HitSpan100" values of the final set of 'low' genomic loci. As noted elsewhere herein, advantageously, the use of HitspanN (e.g. where N is 100) is more successful in visualising loci that are useful to the present invention relative to using sequencing depth. This is because a high Hitspan100 is achieved for a plasma sample only if there is very little or no fragmentation of cfDNA across the sequence which spans a site with at least 50 nucleotides both upstream and downstream. The advantage of using Hitspan100 lies in its ability to identify sites which remain highly intact across 50bp either side of the site or on the other hand, are fragmented 50bp either side of the site.

Next, the ability of the nucleosomal plasma separation score to distinguish between plasma and whole blood samples was tested. Plasma samples originating from different cohorts were tested (i.e., using samples obtained from different clinical sites, diverse ethnicities, different average sample sequencing depths (600X and 50X) and different types of blood collection tubes (EDTA and Streck tubes)):

- Cohort 1: Average sample sequencing depth of 600X, blood collection in EDTA tubes
- Cohort 2: Average sample sequencing depth of 50X, blood collection in EDTA tubes
- Cohort 3: Average sample sequencing depth of 50X, blood collection in Streck tubes

The nucleosomal plasma separation score was consistently able to distinguish between plasma and whole blood samples, with a 100% separation between the scores obtained for whole blood and those

obtained for plasma (Figure 7A). The scores obtained for whole blood were close to 1, whereas the scores obtained for plasma ranged from about 1.2 to about 2 (Figure 7B).

The nucleosomal plasma separation scores were insignificantly different for samples obtained from human subjects that had been diagnosed with lung cancer.

5 **Example 5B: Use of the assay to determine nucleosomal plasma separation in an uncharacterised sample of unknown purity**

An uncharacterised plasma sample is provided, for which the degree of whole blood DNA contamination is unknown.

A nucleosomal plasma separation score is determined using the method as described in Example 5A.

10 Briefly, the score is calculated by the sum of the “HitSpan100” values of the final set of ‘high’ genomic loci in the uncharacterized sample divided by the sum of “HitSpan100” values of the final set of ‘low’ genomic loci in the uncharacterized sample.

The degree of contamination of whole blood DNA in the uncharacterised sample is inferred by rule of proportionality, by comparing its nucleosomal plasma separation score against the reference

15 nucleosomal plasma separation score obtained for pure plasma samples and the nucleosomal plasma separation score obtained for whole blood samples.

CLAIMS

1. A method for determining efficiency of plasma separation from whole blood, comprising:
 - (a) generating from a plasma DNA sample, by PCR amplification, a first amplification product
5 from a first genomic locus using a first primer pair, and a second amplification product from a second genomic locus using a second primer pair, wherein the concentration ratio of the first genomic locus to the second genomic locus in cfDNA is different from the concentration ratio of the first genomic locus to the second genomic locus in whole blood DNA and wherein:
 - 10 i. the lengths (in nucleotides) of the first and second amplification products are within 35% of their mean length; and/or
 - ii. the first and second amplification products are <350 nucleotides in length;
 - (b) quantitating the amplification products; and
 - (c) determining the efficiency of plasma separation based on the quantitated amplification
15 products.
2. The method of claim 1, wherein the concentration ratio of the first locus to the second locus in cfDNA differs from the concentration ratio of the first locus to the second locus in whole blood DNA by a factor of at least 1.5, or a factor of at least 2.
3. The method of any preceding claim, wherein the PCR in step (a) is real-time PCR.
- 20 4. The method of any preceding claim, wherein the method comprises the use of fluorescent probes for specifically detecting the amplification products.
5. The method of any preceding claim, wherein step (b) includes determining a quantification cycle (Cq) for the first and second amplification products, and wherein efficiency of plasma separation in (c) is determined:
 - 25 A. based on a comparison of the Cq value of the first amplification product to a predefined threshold; or
 - B. based on a comparison to a predefined threshold of the difference in Cq (ΔCq) for the first and second amplification products.

6. The method of any preceding claim, wherein the PCR efficiency of generating the first amplification product is equal to the PCR efficiency of generating the second amplification product.
7. The method of any preceding claim, wherein the lengths (in nucleotides) of the first and second amplification products are the same, or within 15% of their mean length.
8. The method of any preceding claim, wherein the first amplification product comprises or consists of:
 - A. the sequence according to any one of SEQ ID NOs: 1 or 2; or
 - B. a sequence with at least about 95% sequence identity, preferably at least 98%, more preferably at least 99% or at least 99.5%, to the nucleotide sequence according to any of SEQ ID NOs: 1 or 2.
9. The method any preceding claim, wherein the second amplification product comprises or consists of the sequence according to SEQ ID NO: 3.
10. The method of any preceding claim, wherein:
 - A. the first and second amplification products comprise sequences according to SEQ ID NO: 1 and SEQ ID NO: 3, respectively; or
 - B. the first and second amplification products comprise sequences according to SEQ ID NO: 2 and SEQ ID NO: 3, respectively.
11. The method of any preceding claim, wherein the plasma sample is a human plasma sample.
12. A method for determining the efficiency of plasma separation from whole blood, comprising:
 - (a) generating from a plasma DNA sample, by PCR amplification, a first amplification product from a first genomic locus using a first primer pair, wherein the first locus is present at different concentrations in cfDNA and whole blood DNA;
 - (b) quantitating the amplification product; and
 - (c) determining if the plasma sample is separated based on the quantitated amplification product.
13. The method of claim 12, wherein the concentration of the first locus in cfDNA differs from the concentration of the first locus in whole blood DNA by a factor of at least 1.5, or a factor of at least 2.

14. The method of claim 12 or 13, further characterised as defined in any of claims 3, 4, 5, 8 and/or 11.
15. A method for analysing a plasma sample, wherein the method comprises a step of determining plasma separation efficiency by the method according to any preceding claim.
- 5 16. A method for preparing a plasma sample, wherein the method comprises a step of determining plasma separation efficiency by the method according to any preceding claim.
17. A method for diagnosis, wherein the method comprises a step of determining plasma separation efficiency by the method according to any preceding claim.
18. A method for detecting disease, wherein the method comprises a step of determining plasma
10 separation efficiency by the method according to any preceding claim.
19. A method of treating a disease in a subject, comprising steps of:
- (a) determining plasma separation efficiency of plasma obtained from the subject by the method according to any preceding claim; and
 - (b) administering a treatment to a subject in need thereof or performing surgery on a
15 subject in need thereof.
20. A method of determining a threshold value, in comparison to which a plasma sample may be determined to be efficiently separated, comprising:
- (i) spiking a pure plasma sample with different amounts of spiking DNA to obtain a series of
20 spiked plasma samples with different degrees of spiking DNA contamination, wherein the spiking DNA is buffy coat DNA or whole blood DNA;
 - (ii) generating for each spiked sample, by PCR amplification, a first amplification product from a first genomic locus using a first primer pair, and a second amplification product from a second genomic locus using a second primer pair, wherein the concentration ratio of the first genomic locus to the second genomic locus in cfDNA is different from the
25 concentration ratio of the first genomic locus to the second genomic locus in whole blood DNA;
 - (iii) calculating a ΔC_q for each amplification product and each spiked sample and determining the minimum or maximum threshold according to a chosen degree of spiking DNA contamination.

21. A method of determining a threshold value, in comparison to which a plasma sample may be determined to be efficiently separated, comprising:
- (i) spiking a pure plasma sample with different amounts of spiking DNA to obtain a series of spiked plasma samples with different degrees of spiking DNA contamination, wherein the spiking DNA is buffy coat DNA or whole blood DNA;
 - (ii) generating for each spiked sample, by PCR amplification, a first amplification product from a first genomic locus using a first primer pair, wherein the first locus is present at different concentrations in cfDNA and whole blood DNA;
 - (iii) calculating a C_q for the first amplification product and for each spiked sample and determining the minimum or maximum threshold according to a chosen degree of spiking DNA contamination.
22. The method of claim 20, wherein:
- i. the lengths (in nucleotides) of the first and second amplification products are within 35% of their mean length; and/or
 - ii. the first and second amplification products are <350 nucleotides in length.
23. The method of any one of claims 20 to 22, wherein the lengths (in nucleotides) of the first and second amplification products are the same, or within 15% of their mean length.
24. A kit for determining the efficiency of plasma separation from whole blood, comprising:
- A. primers for amplifying at least one genomic locus, wherein:
 - i. the primers comprise a first primer pair for amplifying a first genomic locus, wherein the first locus is present at different concentrations in cfDNA and whole blood DNA; or
 - ii. the primers comprise a first primer pair for amplifying a first genomic locus and a second primer pair for amplifying a second genomic locus, wherein the concentration ratio of the first genomic locus to the second genomic locus in cfDNA is different from the concentration ratio of the first genomic locus to the second genomic locus in whole blood DNA and wherein:
 - a. the lengths (in nucleotides) of the first and second amplification products are within 35% of their mean length; and/or
 - b. the first and second amplification products are <350 nucleotides in length;
 - and B. probes for quantitatively detecting the amplification product(s) produced by the primers of A.

25. The kit of claim 24, wherein the lengths (in nucleotides) of the first and second amplification products are the same, or within 15% of their mean length.
26. The kit of any one of claims 24-25, further comprising an instruction manual directing the correlation between quantitated amplification products and level of separation, optionally
5 wherein:
(i) the instruction manual provides a threshold metric, based on which a plasma sample is determined to be separated; or
(ii) the instruction manual provides a threshold C_q, based on which a plasma sample is determined to be separated; or
10 (iii) instruction manual provides a threshold deltaC_q (C_q(second) -C_q(first)), based on which a plasma sample is determined to be separated.
27. The kit of any one of claims 24-26, wherein the probes are fluorescently-labeled oligonucleotide probes.
28. The kit of any one of claims 24-27, wherein the PCR using the first and second primer pairs are
15 of equal efficiency.
29. The kit of any one of claims 24-28, wherein:
A. the first amplification product comprises:
i. the sequence set forth in SEQ ID NO: 1 or a sequence with at least about 95%
sequence identity, preferably at least 98%, more preferably at least 99% or at least
20 99.5%, to the nucleotide sequence of SEQ ID NO:1; or
ii. the sequence set forth in SEQ ID NO: 2 or a sequence with at least about 95%
sequence identity, preferably at least 98%, more preferably at least 99% or at least
99.5%, to the nucleotide sequence of SEQ ID NO:2; and/or
B. the second amplification product comprises the sequence set forth in SEQ ID NO: 3 or a
25 sequence with at least about 95% sequence identity, preferably at least 98%, more preferably at
least 99% or at least 99.5%, to the nucleotide sequence of SEQ ID NO:3.
30. The kit of claim any of claims 24-29, wherein:
A. the first primer pair is:
i. SEQ ID NO: 4 and SEQ ID NO: 5; or
30 ii. SEQ ID NO: 6 and SEQ ID NO: 7; and/or
B. the second primer pair is SEQ ID NO: 8 and SEQ ID NO: 9.

31. A method for determining efficiency of plasma separation from whole blood, comprising:
(a) quantifying, from high-throughput sequencing data of a plasma sample, read counts of a set of first genomic loci and read counts of a set of second genomic loci, wherein the ratio of the sum of the concentrations of the genomic loci of the first set of genomic loci to the sum of the concentrations of the genomic loci of the second set of genomic loci in cfDNA is different from the ratio of the sum of the concentrations of the genomic loci of the first set of genomic loci to the sum of the concentrations of the genomic loci of the second set of genomic loci in whole blood DNA; and
(b) determining the efficiency of plasma separation based on the quantitated read counts.
32. The method of claim 31, wherein step (b) comprises determining the efficiency of plasma separation based on a nucleosomal plasma separation score, wherein the nucleosomal plasma separation score is calculated by dividing the sum of the read counts of the set of first genomic loci by the sum of the read counts of the set of second genomic loci.
33. The method of claim 31 or 32, wherein the read count of a locus is the HitSpanN value of the locus, wherein the HitSpanN value of the locus is number of sequence reads which span the locus with at least $N/2$ nucleotides both upstream and downstream, wherein N is an even number between 2 and 150.
34. The method of claim 33, wherein the read count of a locus is the HitSpan100 value of the locus, wherein the HitSpan100 value of the locus is the number of sequence reads which span the locus with at least 50 nucleotides both upstream and downstream.
35. The method of any one of claims 31-34, wherein the set of first genomic loci consists of 1 genomic locus, or comprises 10 genomic loci, 100 genomic loci, 1,000 genomic loci, 10,000 genomic loci, 100,000 genomic loci or 1,000,000 genomic loci.
36. The method of any one of claims 31-35, wherein the set of second genomic loci consists of 1 genomic locus, or comprises 10 genomic loci, 100 genomic loci, 1,000 genomic loci, 10,000 genomic loci, 100,000 genomic loci or 1,000,000 genomic loci.
37. The method of any one of claims 31-36, wherein the set of first genomic loci and the set of second genomic loci each consist of 1 genomic locus, or comprise 10 genomic loci, 100 genomic loci, 1,000 genomic loci, 10,000 genomic loci, 100,000 genomic loci or 1,000,000 genomic loci.
38. The method of any one of claims 31-34, wherein the set of first genomic loci comprises at least 1,000,000 loci and the set of second genomic loci comprises at least 1,000,000 loci.

39. The method of any one of claims 31-38, wherein the loci of the first and/or the second set of loci are not located on sex chromosomes.
40. The method of any one of claims 31-39, wherein the loci of the first and/or the second set of loci are not located on mitochondrial DNA.
- 5 41. The method of any one of claims 31-40, wherein the loci of the first and/or the second set of loci are located at least 100 bp away from restriction sites of enzymes used for digestion during preparation of the DNA sample used in high-throughput sequencing.
42. The method of any one of claims 31-41, wherein each of the loci of the first and/or the second set of loci corresponds to single positions in the genome.
- 10 43. The method of claim 31-42, wherein the first and/or the second set of loci have a number of sequence reads between 0.5-1.5 of the median number of sequence reads of all the loci in the high-throughput sequencing data.
44. The method of any one of claims 31-43, wherein the plasma sample is a human plasma sample.
45. A method for determining efficiency of plasma separation from whole blood, comprising:
- 15 (a) quantifying, from high-throughput sequencing data of a plasma sample, the read counts of a set of first genomic loci, wherein the sum of the concentrations of the genomic loci of the first set of genomic loci is present at different concentrations in cfDNA and whole blood DNA; and
- (b) determining the efficiency of plasma separation based on the quantitated read counts.
46. The method of claim 45, wherein step (b) comprises normalizing the read counts determined in
- 20 step (a) against the median read count of a set of normalising loci in the high-throughput sequencing data, to obtain normalized read counts and determining the efficiency of plasma separation based on the sum of the normalized read counts.
47. The method of claim 45 or claim 46, further characterised as defined in any of claims 33-35 or 39-44.
- 25 48. The method of any one of claims 45-47, wherein the loci of the set of normalising loci are not located on sex chromosomes.
49. The method of any one of claims 45-48, wherein the loci of the set of normalising loci are not located on mitochondrial DNA.

50. The method of any one of claims 45-49, wherein the loci of the set of normalising loci are located at least 100 bp away from restriction sites of enzymes used for digestion during preparation of the high-throughput sequencing data.
51. The method of any one of claims 45-50, wherein each of the loci of the set of normalising loci corresponds to single positions in the genome.
52. The method of any one of claims 45-51, wherein the set of normalising loci have a number of sequence reads between 0.5-1.5 of the median number of sequence reads of all the loci in the high-throughput sequencing data.
53. The method of any one of claims 45-52, wherein the plasma sample is a human plasma sample.
54. A method for analysing a plasma sample, wherein the method comprises a step of determining plasma separation efficiency by the method according to any of claims 31-53.
55. A method for preparing a plasma sample, wherein the method comprises a step of determining plasma separation efficiency by the method according to any of claims 31-53.
56. A method for diagnosis, wherein the method comprises a step of determining plasma separation efficiency by the method according to any of claims 31-53.
57. A method for detecting disease, wherein the method comprises a step of determining plasma separation efficiency by the method according to any of claims 31-53.
58. A method of treating a disease in a subject, comprising steps of:
- (a) determining plasma separation efficiency of plasma obtained from the subject by the method according to any of claims 31-53; and
 - (b) administering a treatment to a subject in need thereof or performing surgery on a subject in need thereof.
59. A method of providing a parameter, comprising:
- (a) from sequencing data, defining one or more genomic regions;
 - (b) for each genomic region, determining a local maximum within the genomic region as the parameter, the local maximum being the locus having the highest read count within the genomic region.
60. A method of providing a parameter, comprising:

- (a) from sequencing data, defining one or more genomic regions;
- (b) for each genomic region, determining a local minimum within the genomic region as the parameter, the local minimum being the locus having the lowest read count within the genomic region.

- 5 61. The method of claim 59 or claim 60, wherein the sequencing data are from plasma samples.
62. The method of any one of claims 59-61, wherein step (a) comprises defining a set of consecutive 200-bp genomic regions from sequencing data.
63. A method of locating nucleosome sites, comprising the method of any one of claims 59-62.
64. A method of determining two sets of loci, using which a plasma sample may be determined to be efficiently separated, comprising:
- 10 (i) obtaining high-throughput sequencing data of a pure plasma sample;
- (ii) determining a set of loci using the method of providing a parameter of any one of claims 59-63, wherein the parameter is a maximum; and determining a set of loci using the method of providing a parameter of any one of claims 59-63, wherein the parameter is a minimum.
- 15 65. A method of determining a set of loci, using which a plasma sample may be determined to be efficiently separated, comprising:
- (i) obtaining high-throughput sequencing data of a pure plasma sample;
- (ii) determining a set of loci using the method of any one of claims 59-63.
- 20 66. A method of determining a threshold value, in comparison to which a plasma sample may be determined to be efficiently separated, comprising:
- (i) spiking a pure plasma sample with different amounts of spiking DNA to obtain a series of spiked plasma samples with different degrees of spiking DNA contamination, wherein the spiking DNA is buffy coat DNA or whole blood DNA;
- (ii) generating for the pure plasma sample and each spiked sample, high-throughput sequencing data;
- 25 (iii) for each spiked sample, determining the read counts of a set of first genomic loci and the read counts of a set of second genomic loci, wherein the ratio of the sum of the concentrations of the genomic loci of the first set of genomic loci to the sum of the concentrations of the genomic loci of the second set of genomic loci in cfDNA is different from the ratio of the sum of the concentrations of the genomic loci of the first set of genomic loci to the sum of the concentrations of the genomic loci of the second
- 30

set of genomic loci in whole blood DNA, and determining the minimum or maximum threshold value according to a chosen degree of spiking DNA contamination.

67. A method of determining a threshold value, in comparison to which a plasma sample may be determined to be efficiently separated, comprising:
- (i) spiking a pure plasma sample with different amounts of spiking DNA to obtain a series of spiked plasma samples with different degrees of spiking DNA contamination, wherein the spiking DNA is buffy coat DNA or whole blood DNA;
 - (ii) generating for the pure plasma sample and each spiked sample, high-throughput sequencing data;
 - (iii) for each spiked sample, determining the read counts of a set of first genomic loci, wherein the sum of the concentrations of the genomic loci of the first set of genomic loci is present at different concentrations in cfDNA and whole blood DNA, and determining the minimum or maximum threshold according to a chosen degree of spiking DNA contamination.