

DETECTION OF VIRAL PRESENCE IN CELLS**FIELD OF THE INVENTION**

The present invention relates to a method that can detect the presence or absence of a virus
5 and/or promoter thereof and the possible activity of the virus in the cell. In particular, the method quantitatively measures the level of DNA methylation on the CpG sites of the viral DNA in the cells.

BACKGROUND OF THE INVENTION

Viral infection results in a variety of human diseases including cancer. Upon infecting host cells, the
10 viruses hijack cellular machinery to promote their replication and produce viral particle progeny. To promote the long-term persistence of the virus in the cell, some viruses integrate their genomic DNA into the host genome which results in several consequences such as gene disruption, chromosome instability, gene mutation and oncogenesis. Such viral DNA which is integrated into the host genome is known as provirus. A provirus passively gets replicated along with the host
15 genome; thus, it gets passed down from one generation to the next. Integration of viral DNA into the host genome may lead to a latent infection or a productive infection.

In case of a productive infection, the new virus particles will be produced via transcription of the provirus to generate mRNA and thus spreading infection to other cells. Latent infection on the other
20 hand, refers to a state where the provirus is transcriptionally silent, however, the provirus gets activated to start transcription in response to changes in the host's environmental conditions. Hence, the host cell's protein synthesis machinery is hijacked to generate more viruses leading to the destruction of the host cell.

25 Accordingly, it is necessary not to only check the presence or absence of viral DNA in a host cell but also predict the viral activity in the host cell to determine the seriousness of the viral presence. Further, in cell line assessment, vaccine development, patient samples etc as a part of quality control, it is an absolute necessity to check for viral DNA integration.

30 Viral vectors are prevalently used to express transgene for gene therapy, cell line development, and therapeutic protein production. A viral vector is designed by substituting a set of the viral gene responsible for viral replication and pathogenicity with the transgene cassette. Gene transfer using viral vectors results in a high and long-term expression of the transgene. The viral promoter and enhancer are the major DNA regulatory elements in the viral vector that determines the transgene
35 expression level in the cells. The integration sites of the viral vector are susceptible to transcriptional regulation via epigenetic regulation such as histone modifications and DNA methylation. For example, the DNA methylation status of the viral vector is an important factor for protein production or expression stability in producer cells. An increase in DNA methylation on the CpG sites of the viral promoter results in transgene silencing at transcription levels. For example,
40 the protein production variability in CHO cells has been associated with DNA methylation mediated regulation of Cytomegalovirus Major Immediate-Early and enhancer (CMV) promoter and simian

vacuolating virus 40 (SV40) promoters which are the most frequently used promoters for the production of recombinant proteins in CHO cells.

5 However, as discussed above, the presence of viral DNA in a host cell is not representative of the host cell being in serious danger or being infected by the virus. The current methods used to test the presence or integration of viral DNA in a host cell, lack the ability to determine the potency of the presence of the virus on the survival of the host cell. Accordingly, there is a need in the art that quantitatively and qualitatively not only determines the presence of a virus in a host cell but also the potency of the DNA integration on the host cell.

10

DESCRIPTION OF THE INVENTION

The present invention solves the problems above by providing a DNA-based method for determining the level of DNA methylation on the CpG sites on the viral DNA and/or viral promoter which can then be used to not only predict the viral activity on the host cell but also the expression of a transgene in cells. In particular, the method according to any aspect of the present invention may be carried out on a DNA-based array. The method according to any aspect of the present invention may thus provide an efficient, cheap and accurate means of determining not only the presence of a virus in a cell but also the activity of the virus in the cell.

20 According to one aspect of the present invention, there is provided a method of identifying integrated viral DNA in the genome of at least one mammalian test cell, the method comprising the steps of:

- (a) determining presence or absence of viral DNA in a DNA sample obtained from the mammalian test cell; and
- 25 (b) determining the methylation status of at least one CpG site in the viral DNA of the DNA sample obtained from the mammalian test cell,

wherein the presence of the viral DNA after step (a) and hypermethylation in step (b) is indicative of the viral DNA being integrated in the mammalian test cell and the virus being inactive respectively; the presence of the viral DNA after step (a) and baseline methylation or hypomethylation in step (b) is indicative of the viral DNA being integrated in the mammalian test cell and the virus being active respectively; and

30 the absence of the viral DNA after step (a) is indicative of the viral DNA not being integrated in the mammalian test cell respectively; and

wherein the method is carried out on an array.

35

The term "viral DNA integration" used synonymously with "integrated viral DNA" or "genome integrated viral DNA" refers to the incorporation of the whole genome or part of the genome of at least one virus into a host cell, particularly an animal cell, more particularly a mammalian cell. Viral DNA integration is a means of viral insertion that is not transient as the viral DNA is genomically integrated into the host cell, i.e. mammalian cell. Viral DNA integration is a potential diagnostic and prognostic marker. In particular, viral DNA integration is a unique enzymatic process where double-stranded linear viral DNA is inserted into the host genome in a catalysed by a virus-encoded

40

integrase. Many of these viruses however lie dormant in the infected cell, although their DNA may be integrated into the DNA of the host cell chromosome. Some viruses, present in the host cells may also lead to the cell being malignant and eventually causing tumours. The family of *Polyomaviruses*, e.g. simian virus 40 (SV40), and viruses from the *Retroviridae* family and *Herpesviridae* family are examples of viruses that can also be associated with malignancy (causing death or illness) in the host cell. The viral DNA may be transformed from the viral RNA.

A "mammalian cell" as used herein refers to is a cell from any member of the order Mammalia which includes a cell from a mouse, a rat, a monkey, a guinea pig, a dog, a mini-pig, a human being, a cow, a sheep, a pig, a goat, a horse, a donkey, a mule, a hamster or the like. The mammalian cell may also include an established cell line or immortalized cell line. For example, the mammalian cell according to any aspect of the present invention may be a CHO cell line which refers to immortal Chinese Hamster Ovary cell line (CHO) derived from *Cricetulus griseus*. In particular, the CHO cell line may be selected from the group consisting of CHO-K1 (ATCC), CHO-DG44 (Thermo Fisher Scientific), CHO-DXB11 (ATCC), ExpiCHO-S™ cells (Thermo Fisher Scientific), FreeStyle™ CHO-S™ cells (Thermo Fisher Scientific), CHO 1-15 [subscript 500] (ATCC), Agarabi CHO (ATCC), and a CHOK1SV cell including all variants (e.g. POTELLIGENT®, Lonza, Slough, UK), a CHOK1SV GS-KO (glutamine synthetase knockout) cell including all variants (e.g., XCEED™ Lonza, Slough, UK). The mammalian cell may be from Baby Hamster Kidney fibroblasts (BHK (ATCC CCL-10), or Vero cell (ATCC CCL-81). Exemplary human cells include human embryonic kidney (HEK) cells, such as HEK293 (ATCC CRL-1573), HEK 293T (ATCC CRL-3216), a HeLa cell (ATCC CCL-2), a NS0 cell (ECACC 85110503), or a Sp2/0 cell (ATCC CRL-1581). The mammalian cells according to any aspect of the present invention may include mammalian cell cultures which can be either adherent cultures or suspension cultures.

According to any aspect of the present invention, the first step involves:

- (a) determining presence or absence of viral DNA in a DNA sample obtained from the mammalian test cell.

The presence or absence of a viral DNA in a DNA sample may be determined by establishing a viral titre in a biological sample, i.e. DNA from the test mammalian cell. As used herein "viral titre" refers to a numeric expression of the quantity of a virus in a given volume, generally expressed as viral particles, transducing units, or infections particles, per millilitre (mL).

The presence or absence of viral DNA in a DNA sample obtained from the mammalian test cell may be determined using a DNA-based array according to any aspect of the present invention. The DNA-based array according to any aspect of the present invention may then comprise probes that bind to any region (i.e. CpG and non-CpG sites) that is specific to the viral DNA. When a probe specific to at least one virus CpG or non CpG site binds to the DNA test sample, that is an indication of the presence of the viral DNA and/or the presence of the viral DNA being integrated into the DNA of the mammalian test cell. More in particular, the array according to any aspect of

the present invention comprises probes that are specific to non-CpG sites that can detect the presence of integrated viral DNA in the genome of the DNA sample to be tested.

5 The presence of specific CpG and non-CpG sites in step (a) is indicative of the viral strain or variant thereof. The term 'variant' as used herein in respect to a virus refers to a virus strain that has changed most likely through mutation to a version that is different from the original virus. The term "variant" refers to a form of the virus strain that deviates from what occurs in nature or which is considered the wild type. The variant may be 50, 55, 60, 65, 70, 75, 80, 85, 90, 95% similar in sequence compared to the wild-type strain.

10

The second step according to any aspect of the present invention involves:

(b) determining the methylation status of at least one CpG site in the viral DNA of the DNA sample obtained from the mammalian test cell.

15 If in step (a) the presence and integration of the viral DNA or integrated viral DNA in the mammalian DNA is confirmed, then the methylation status of at least one of the CpGs of the integrated viral DNA may be determined from the DNA of the mammalian test cell.

20 As used herein, a "CpG site" or "methylation site" is a nucleotide within a nucleic acid (DNA or RNA) that is susceptible to methylation either by natural occurring events *in vivo* or by an event instituted to chemically methylate the nucleotide *in vitro*. Some of these sites may be hypermethylated and some may be hypomethylated in a cell.

25 As used herein, a "methylated nucleic acid molecule" refers to a nucleic acid molecule that contains one or more nucleotides that is/are methylated.

30 As used herein, a "methylated nucleotide" or a "methylated nucleotide base" refers to the presence of a methyl moiety on a nucleotide base, where the methyl moiety is usually not present in a recognized typical nucleotide base. For example, cytosine in its usual form does not contain a methyl moiety on its pyrimidine ring, but 5-methylcytosine contains a methyl moiety at position 5 of its pyrimidine ring. Therefore, cytosine in its usual form may not be considered a methylated nucleotide and 5-methylcytosine may be considered a methylated nucleotide. In another example, thymine may contain a methyl moiety at position 5 of its pyrimidine ring, however, for purposes herein, thymine may not be considered a methylated nucleotide when present in DNA. Typical
35 nucleotide bases for DNA are thymine, adenine, cytosine and guanine. Typical bases for RNA are uracil, adenine, cytosine and guanine. Correspondingly a "methylation site" is the location in the target gene nucleic acid region where methylation has the possibility of occurring. For example, a location containing CpG is a methylation site wherein the cytosine may or may not be methylated. In particular, the term "methylated nucleotide" refers to nucleotides that carry a methyl group
40 attached to a position of a nucleotide that is accessible for methylation. These methylated nucleotides are usually found in nature and to date, methylated cytosine that occurs mostly in the context of the dinucleotide CpG, but also in the context of CpNpG- and CpNpN-sequences may be

considered the most common. In principle, other naturally occurring nucleotides may also be methylated but they will not be taken into consideration with regard to any aspect of the present invention.

- 5 A "CpG island" as used herein describes a segment of DNA sequence that comprises a functionally or structurally deviated CpG density. For example, Yamada et al. have described a set of standards for determining a CpG island: it must be at least 400 nucleotides in length, has a greater than 50% GC content, and an OCF/ECF ratio greater than 0.6 (Yamada et al., 2004, Genome Research, 14, 247-266). Others have defined a CpG island less stringently as a sequence at least
- 10 200 nucleotides in length, having a greater than 50% GC content, and an OCF/ECF ratio greater than 0.6 (Takai et al., 2002, Proc. Natl. Acad. Sci. USA, 99, 3740-3745). In context of the present invention, the terms "methylation profile", "methylation pattern", "methylation state" or "methylation status," are used herein to describe the state, situation or condition of methylation of a genomic sequence, and such terms refer to the characteristics of a DNA segment at a particular genomic
- 15 locus in relation to methylation. Such characteristics include, but are not limited to, whether any of the cytosine (C) residues within this DNA sequence are methylated, location of methylated C residue(s), percentage of methylated C at any particular stretch of residues, and allelic differences in methylation due to, e.g., difference in the origin of the alleles.
- 20 The term "methylation status" refers to the status of a specific methylation site (i.e. methylated vs. non-methylated) which means a residue or methylation site is methylated or not methylated. Then, based on the methylation status of one or more methylation sites, a methylation profile may be determined.
- 25 The term "methylation level" refers to the level of a specific methylation site which can range from 0 (=unmethylated) to 1 (= fully methylated). Thus, based on the methylation level of one or more methylation sites, a methylation profile may be determined. Accordingly, the term "methylation profile" or also "methylation pattern" refers to the relative or absolute concentration of methylated C or unmethylated C at any particular stretch of residues in a biological sample. For example, if
- 30 cytosine (C) residue(s) not typically methylated within a DNA sequence are more methylated in a sample, it may be referred to as "hypermethylated"; whereas if cytosine (C) residue(s) typically methylated within a DNA sequence are less methylated, it may be referred to as "hypomethylated". Likewise, if the cytosine (C) residue(s) within a DNA sequence (e.g., sample nucleic acid) are more methylated when compared to another sequence from a different region or from a different
- 35 individual (e.g., relative to normal nucleic acid), that sequence is considered hypermethylated compared to the other sequence. Alternatively, if the cytosine (C) residue(s) within a DNA sequence are less methylated as compared to another sequence from a different region or from a different individual, that sequence is considered hypomethylated compared to the other sequence. These sequences are said to be "differentially methylated". For example, when the methylation
- 40 status differs between inflamed and non-inflamed tissues, the sequences are considered "differentially methylated". Measurement of the levels of differential methylation may be done by a variety of ways known to those skilled in the art. One method is to measure the methylation level of

individual interrogated CpG sites determined by the bisulfite sequencing method, as a non-limiting example.

- Bisulfite treatment' of genomic DNA used interchangeably with the term 'bisulfite modification', refers to the treatment of the genomic DNA with a deaminating agent such as a bisulfite that may be used to treat all DNA, methylated or not. In particular, the term "bisulfite" as used herein encompasses any suitable type of bisulfite, such as sodium bisulfite, or other chemical agents that are capable of chemically converting a cytosine (C) to an uracil (U) without chemically modifying a methylated cytosine and therefore can be used to differentially modify a DNA sequence based on the methylation status of the DNA, e.g., U.S. Pat. Pub. US 2010/0112595. As used herein, a reagent that "differentially modifies" methylated or non-methylated DNA encompasses any reagent that modifies methylated and/or unmethylated DNA in a process through which distinguishable products result from methylated and non-methylated DNA, thereby allowing the identification of the DNA methylation status. Such processes may include, but are not limited to, chemical reactions (such as a C to U conversion by bisulfite) and enzymatic treatment (such as cleavage by a methylation-dependent endonuclease). Thus, an enzyme that preferentially cleaves or digests methylated DNA is one capable of cleaving or digesting a DNA molecule at a much higher efficiency when the DNA is methylated, whereas an enzyme that preferentially cleaves or digests unmethylated DNA exhibits a significantly higher efficiency when the DNA is not methylated.
- Accordingly, before step (a) according to any aspect of the present invention is carried out, the genomic DNA contained/ obtained or extracted from the cell, is first bisulfite treated.

- An alternative method available in the art may be used instead of bisulfite treatment. A skilled person will understand which other methods to use. In one example, TET-assisted pyridine borane sequencing (TAPS) may be used for detection of 5mC and 5hmC (Yibin Liu, et al., *Nature Biotechnology*, 37: 424–429 (2019)).

- As used herein, the term "genomic material" refers to nucleic acid molecules or fragments of the genome of the animal according to any aspect of the present invention. In particular, such nucleic acid molecules or fragments are DNA or RNA or hybrids thereof, and most preferably are molecules of the DNA genome of a subject or group of subjects.

- The term 'biological sample' as used herein may be selected from the group consisting of muscle, organ tissue, milk, blood, brain, sperm and any other tissue or sample that provides genomic DNA to be used in the method according to any aspect of the present invention. In particular, the biological sample may comprise any biological material obtained from the subject that contains DNA, and may be liquid, solid or both, may be tissue or bone, or a body fluid such as blood, lymph, etc. In particular, the biological sample useful for the present invention may comprise biological cells or fragments thereof.

- As used herein, the "DNA sample" refers to the DNA extracted from a cell of the mammal according to any aspect of the present invention using known methods in the art. In particular, in

order to detect the presence or absence of integrated viral DNA in the genome of the DNA tested, the DNA used for the testing is purified high-molecular-weight genomic fragment of size ranging from 50 to 150 kb. Further washing steps removes low molecular weight DNA fragments which represents episomal DNA, virion DNA and cccDNA.

5

The term "test" used in conjunction with the term mammal or mammalian cell herein refers to the mammal or cell that is to be introduced to the array according to any aspect of the present invention and is the basis for an analysis application of the present invention to determine if the cell has viral DNA integrated in the test cell genome and whether this virus is active or not. A "test profile" is therefore from an (individual) subject or group of subjects being tested according to the invention or a profile being obtained or generated in this context. Similarly, the term 'sample' used in accordance with any aspect of the present invention refers to an entity that may be subject to the method of the present invention. Conversely, the term "reference" or 'control' shall denote, mostly predetermined, entities which are used for a comparison with the test entity. In particular, a sample may be extracted DNA from a test mammalian cell that may be subject to the method of the present invention to determine if there is viral DNA integrated in the mammalian cell and if the virus is active in the test cell by first determining the presence or absence of the virus in the cell and then determining the DNA methylation profile of the test cell. There may be a further step of comparing this test methylation profile with a control and a 'control' refers to a mammalian cell, possibly of the same taxon, species, or cell line where the features as mentioned above are already known and where the methylation status is already known and used as a reference.

The DNA methylation profile of step (a) and (b) according to any aspect of the present invention is determined using an array. In particular, a bead-based array. The array according to any aspect of the present invention is advantageous as it not only enables early detection of integrated viral DNA in a host mammalian cell but also helps in determining the effectiveness of a virus vector in a cell to be used for heterologous protein production. In particular, the method according to any aspect of the present invention may be used for a better understanding of genome stability of mammalian cell lines and enables better control over the manufacturing/ process development/ product development/ scaling up/ validation process, thereby aiding in the selection of better mammalian cell lines for industrial applications.

Arrays allow for a high-throughput and robust method to determine semi-quantitative/quantitative DNA-methylation information through a small sample of extracted DNA of interest. These custom designed arrays may use Illumina iScan and Infinium platform technology or an equivalent thereof, which allows on each chip for example 100,000 different bead types that covalently bind DNA-methylation probes. Each probe represents one CpG Methylation site at the end of the probe sequence. DNA samples undergo bisulfite conversion, amplification, fragmentation, precipitation and resuspension steps before hybridization on an array chip. Once on the chip the DNA hybridizes to the beads for each CpG site so that methylation changes at each site can be detected specifically through single nucleotide extension. This is especially advantageous as the array-based method is simple and the results of the array are accurate and reproducible.

Further, compared to traditional sequencing which can take weeks to generate data, the array technology has a much shorter turn-around time. The volume and complexity of data generated is lesser compared to sequencing making it computationally less intensive. This allows for quicker
5 computation to achieve interpretable results from experimental groups. Overall microarray technology is roughly 10x faster and 10x cheaper than traditional sequencing while still quantifiable for the methylation level at specific CpG sites.

The term "array" as used herein refers to an intentionally created collection of probe molecules
10 which can be prepared either synthetically or biosynthetically. The probe molecules in the array can be identical or different from each other. The array can assume a variety of formats, for example, libraries of soluble molecules; libraries of compounds tethered to resin beads, silica chips, or other solid supports.

15 In particular, an array provides a convenient platform for simultaneous analysis of large numbers of CpG sites, for example, at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 50, 100, 500, 1000, 5000, 10,000, 100,000 or more sites or loci. In particular, the array comprises a plurality of different probe molecules that can be attached to a substrate or otherwise spatially distinguished in an array. Examples of arrays that may be used according to any aspect of the present invention include slide arrays, silicon
20 wafer arrays, liquid arrays, bead-based arrays and the like. In one example, array technology used according to any aspect of the present invention combines a miniaturized array platform, a high level of assay multiplexing, and scalable automation for sample handling and data processing. In particular, the array according to any aspect of the present invention may be an array of arrays, also referred to as a composite array, having a plurality of individual arrays that is configured to
25 allow processing of multiple samples simultaneously. Examples of composite arrays and the technology behind them are disclosed at least in US 6,429,027 and US 2002/0102578. A substrate of a composite array may include a plurality of individual array locations, each having a plurality of probes, and each physically separated from other assay locations on the same substrate such that a fluid contacting one array location is prevented from contacting another array location. Each array
30 location can have a plurality of different probe molecules that are directly attached to the substrate or that are attached to the substrate via rigid particles in wells (also referred to herein as beads in wells).

In one example, an array substrate can be a fibre optical bundle or array of bundles as described in
35 US6,023,540, US6,200,737 and/or US6,327,410. An optical fibre bundle or array of bundles can have probes attached directly to the fibres or via beads. A skilled person would be able to easily determine which substrate will be most suitable for the array according to any aspect of the present invention. WO2004110246 further discloses other substrates and methods of attaching beads to the substrates that may be used in the array according to any aspect of the present invention.

40

In one example, a surface of the substrate may have physical alterations to enable the attachment of probes or produce array locations. For example, the surface of a substrate can be modified to

contain chemically modified sites that are useful for attaching, either-covalently or non-covalently, probe molecules or particles having attached probe molecules. Probes may be attached using any of a variety of methods known in the art including, an ink-jet printing method, a spotting technique, a photolithographic synthesis method, or printing method utilizing a mask. WO2004110246
5 discloses these techniques in more detail.

In one example, the array according to any aspect of the present invention may be a bead-based array, where the beads are associated with a solid support such as those commercially available from Illumina, Inc. (San Diego, Calif.). An array of beads useful according to any aspect of the
10 present invention can also be in a fluid format such as a fluid stream of a flow cytometer or similar device. Commercially available fluid formats for distinguishing beads include, for example, those used in XMAP(TM) technologies from Luminex or MPSS(TM) methods from Lynx Therapeutics.

The term "solid support", "support", and "substrate" as used herein are used interchangeably and
15 refer to a material or group of materials having a rigid or semi-rigid surface or surfaces. In many examples, at least one surface of the solid support will be substantially flat, although in some examples it may be desirable to physically separate synthesis regions for different compounds with, for example, wells, raised regions, pins, etched trenches, or the like.

20 The array or microarray according to any aspect of the present invention may be a very high-density array, for example, those having from about 10,000,000 probes/cm² to about 2,000,000,000 probes/cm² or from about 100,000,000 probes/cm² to about 1,000,000,000 probes/cm². High density arrays are especially useful according to any aspect of the present invention for including the multitude of CpG sites on the array.

25 The array according to any aspect of the present invention may be used to analyse or evaluate such pluralities of loci simultaneously or sequentially as desired. In one example, a plurality of different probe molecules can be attached to a substrate or otherwise spatially distinguished in an array. Each probe is typically specific for a particular locus and can be used to distinguish
30 methylation state of the locus.

The term "probe molecules" or 'probes' as used interchangeably herein refers to a surface-immobilized molecule that can be recognized by a particular target. Probes used in the array can be specific for the methylated allele of a CpG site, the non-methylated allele of the CpG site or both
35 or for the methylated allele of a non-CpG site, the non-methylated allele of the non-CpG site or both.

The term "target" as used herein refers to a molecule that has an affinity for a given probe molecule. Targets may be naturally occurring or man-made molecules. Also, they can be employed
40 in their unaltered state or as aggregates. Targets may be attached, covalently or noncovalently, to a binding member, either directly or via a specific binding substance. Examples of targets which can be employed according to any aspect of the present invention are methylated and non-

5 methylated CpG sites. Targets are sometimes referred to in the art as anti-probes. As the term targets is used herein, no difference in meaning is intended.

10 The term “complementary” as used herein refers to the hybridization or base pairing between nucleotides or nucleic acids, such as, for instance, between the two strands of a double stranded DNA molecule or between an oligonucleotide primer and a primer binding site on a single stranded nucleic acid to be sequenced or amplified. Complementary nucleotides are, generally, A and T (or A and U), or C and G. Two single stranded RNA or DNA molecules are said to be complementary when the nucleotides of one strand, optimally aligned and compared and with appropriate
15 nucleotide insertions or deletions, pair with at least about 80% of the nucleotides of the other strand, usually at least about 90% to 95%, and more preferably from about 98 to 100%. Perfectly complementary refers to 100% complementarity over the length of a sequence. For example, a 25-base probe is perfectly complementary to a target when all 25 bases of the probe are complementary to a contiguous 25 base sequence of the target with no mismatches between the
20 probe and the target over the length of the probe.

25 The presence of the viral DNA after step (a) according to any aspect of the present invention and hypermethylation in step (b) is indicative of the viral DNA being integrated in the mammalian test cell and the virus being inactive respectively; and
30 the presence of the viral DNA after step (a) and baseline methylation or hypomethylation in step (b) is indicative of the viral DNA being integrated in the mammalian test cell and the virus being active respectively; and
the absence of the viral DNA after step (a) is indicative of the viral DNA not being integrated in the mammalian test cell respectively.

35 The term “hypermethylation” refers to the average methylation state corresponding to an increased presence of 5-mCyt at one or a plurality of CpG dinucleotides within a DNA sequence of a test DNA sample, relative to the amount of 5-mCyt found at corresponding CpG dinucleotides within a normal control DNA sample.

40 The term “hypomethylation” refers to the average methylation state corresponding to a decreased presence of 5-mCyt at one or a plurality of CpG dinucleotides within a DNA sequence of a test DNA sample, relative to the amount of 5-mCyt found at corresponding CpG dinucleotides within a normal control DNA sample.

45 A virus is considered to be in an active state when the virus is not only integrated in the host mammalian cell but also capable of replication and protein production. In order to promote the long-term persistence of the virus in the cell, some viruses integrate their genomic DNA into the host genome which results in several consequences such as gene disruption, chromosome instability,
50 gene mutation and oncogenesis. Such viral DNA which is integrated into the host genome is known as provirus. A provirus passively gets replicated along with the host genome, thus it gets passed from one generation to the next. Integrated viral DNA might lead to a latent infection or a productive

- infection. In case of productive infection, the new virus particles will be produced via transcription of the provirus to generate mRNA and thus infect other cells. 'Latent infection' as used herein refers to a state where the provirus is transcriptionally silent, however, the provirus gets activated to start transcription in response to changes in the host's environmental conditions. Hence, the host cell's protein synthesis machinery is hijacked to generate more viruses leading to the destruction of its host cell. Thus, it is an absolute necessity to check for viral DNA integration or integrated viral DNA for cell line assessment, vaccine development, patient samples etc. as a part of quality control. The converse is a true for a virus in an non-active state.
- 10 Viral promoters are rich in CpG sites which make them more prone to DNA methylation and thus suppressing the protein expression. Accordingly, studying DNA methylation of viruses and their promoters can be used to accurately determine the presence or absence of viral integration and activity of the integrated virus when present in the DNA.
- 15 As used herein, the terms "promoter" or "gene promoter" used interchangeably with the terms 'regulatory region' or 'regulatory sequence' refers to the respective contiguous gene DNA sequence extending from 1.5 kb upstream to 1.5 kb downstream relative to the transcription start site (TSS), or contiguous portions thereof. In particular, 'regulatory region' refers to the respective contiguous gene DNA sequence extending from 1.5 kb upstream to 0.5 kb downstream relative to the TSS. In some examples, 'regulatory region' refers to the respective contiguous gene DNA sequence extending from 1.5 kb upstream to the downstream edge of a CpG island that overlaps with the region from 1.5 kb upstream to 1.5 kb downstream from TSS (and in such cases, may thus extend even further beyond 1.5 kb downstream), and contiguous portions thereof. Change in DNA methylation on the gene promoters responsible for protein glycosylation can lead to an improvement of protein quality. Protein glycosylation is a critical quality attribute that modulate the efficacy, stability, and half-life of a therapeutic protein. It is desirable to obtain a consistent glycoform profile in protein production due to regulatory concerns. Hence, DNA methylation can act as a tool to determine integrated viral DNA in a mammalian host cell.
- 25
- 30 According to another aspect of the present invention, there is provided a method of identifying integrated viral DNA in the genome of at least one mammalian test cell, the method comprising the steps of:
- (a) determining presence or absence of one or more pre-selected CpG and non-CpG sites of the viral DNA in a DNA sample obtained from the mammalian test cell; and/or
 - 35 (b) determining a test methylation profile of the pre-selected CpG and non-CpG sites within the DNA sample obtained from the mammalian test cell; and
- wherein the presence of one or more pre-selected CpG and non-CpG sites after step (a) and hypermethylation in (b) is indicative of the viral DNA being integrated in the mammalian test cell but the virus being inactive respectively;
- 40 the presence of one or more pre-selected CpG and non-CpG sites after step (a) and baseline methylation or hypomethylation is indicative of the viral DNA being integrated in the mammalian test cell and the virus being active respectively; and

the absence of the one or more pre-selected CpG and non-CpG sites after step (a) is indicative of the viral DNA not being integrated in the mammalian test cell respectively; and wherein the method is carried out on an array.

- 5 As used herein, the term "pre-selected sites" used interchangeably with "pre-selected CpG and non-CpG sites" refers to sites that were selected from genes or regions that showed the highest degree of methylation variation during the training of the method and fulfils certain quality criteria such as a minimum sequencing coverage of $\geq 5x$ were considered and for ≥ 5 qualified CpG sites. Additionally, genes that have an average methylation level < 0.1 or an average methylation level
- 10 > 0.9 can be excluded due to their limited dynamic range. "Reference methylation profiles" may be defined on the basis of multiple training samples using multivariate statistical methods, such as such as Principal Component analysis or Multi-Dimensional Scaling.

According to any aspect of the present invention, there may be a further step:

- 15 (c) comparing the methylation status of the CpG site from (b) with that of a mammalian control cell where the viral DNA is integrated within the control cell.

Step (c) is an optional as step (b) of measuring methylation status is sufficient to determine the presence of hypomethylation or hypermethylation of a CpG site which can provide information on the virus and protein production without doing a further comparison in step (c).

20

According to a further aspect of the present invention, there is provided a DNA array-based method of identifying integrated viral DNA in at least one mammalian test cell, the method comprising the steps of:

- (a) determining presence of at least one viral CpG site and/or non-CpG site in a DNA
- 25 sample obtained from the mammalian test cell; and
- (b) determining the methylation status of multiple CpG sites and/or non-CpG sites in the DNA sample obtained from the mammalian test cell, and
- (c) comparing the methylation status of the CpG site and/or non-CpG site from (b) with that of a mammalian control cell where the viral DNA is incorporated within the control cell,
- 30 wherein the presence of the CpG site and/or non-CpG site after step (a) and hypermethylation in (c) is indicative of the viral DNA being incorporated in the mammalian test cell but the virus being inactive respectively;
- the presence of the CpG site and/or non-CpG site after step (a) and baseline methylation or hypomethylation is indicative of the viral DNA being incorporated in the mammalian test cell and
- 35 the virus being active respectively; and
- the absence of the CpG site and/or non-CpG site after step (a) is indicative of the viral DNA not being incorporated in the mammalian test cell respectively.

The method according to any aspect of the present invention comprising the step of:

- (ii) contacting the DNA sample from the test cell with a DNA-based array.

40

The method according to any aspect of the present invention further comprising the step of:

- (i) performing bisulfite modification to the DNA sample before step (a).

The term 'contacting', as used herein, means bringing about direct contact between the genomic material sample and DNA-based array. For example, the genomic material sample may be DNA that is extracted from the biological sample from the test mammal, and this directly brought into contact with the probe in the DNA-based array.

The method according to any aspect of the present invention, wherein the mammalian cell is a CHO cell line and there is a further step (c):

- (c) comparing the methylation status of the CpG site and/or non-CpG site in the viral DNA from (b) with that of:
 - (i) a first CHO control cell where the viral DNA is integrated in the genome of the control cell and the first CHO control cell has optimal heterologous protein production; and/or
 - (ii) a second CHO control cell where the viral DNA is integrated in the genome of the control cell and the first CHO control cell has no optimal heterologous protein production;
- wherein significant similarity of the test methylation status of the CpG site with the first CHO control cell is indicative of the test CHO cell having optimal heterologous protein production, or with the second CHO control cell is indicative of the test CHO cell not having optimal heterologous protein production,
- wherein difference of the test methylation status of the CpG site with the first CHO control cell is indicative of the test CHO cell not having optimal heterologous protein production, or with the second CHO control cell is indicative of the test CHO cell having optimal heterologous protein production.

The term 'heterologous protein production' as used herein refers to the production of a protein which is not endogenous to the cell. It means an expression of a gene or part of a gene, particularly a transgene in a host CHO cell which does not naturally express this gene. The assays that are commonly used to quantify heterologous protein production include enzyme-linked immunosorbent assay (ELISA), chromatography & bioprocess analyser. The term 'host cell' as used herein refers to a cellular system for the expression of heterologous protein. For example, CHO cells are the main hosts for the production of various therapeutic proteins.

As used herein, the term 'therapeutic protein' refers to genetically engineered versions of naturally occurring human proteins. Examples of therapeutic proteins include antibody-based drugs, anticoagulants, blood factors, bone morphogenetic proteins, engineered protein scaffolds, enzymes, growth factors, hormones, interferons, interleukins and the like.

The term 'optimal heterologous protein production' herein refers to CHO cells that are capable of high-level protein production, particularly during industrial production or large-scale production of recombinant proteins, where the protein is usually a functional protein that is not naturally occurring in the wild-type CHO cell. In particular, for optimal heterologous protein production a CHO cell line has minimized metabolic burdens and toxic effects to the cell. More in particular, 'optimal heterologous protein production' refers to high level protein production where the CHO cell line not

only produces a high yield of the protein of interest but also that the protein production is constantly maintained over the period of production (i.e., the prolonged period of culture) such that the quality of the protein produced is also consistent and maintained. In particular, for a CHO cell according to any aspect of the present invention to be capable of 'optimal heterologous protein production', the cell must at least display one of more of the following phenotypes of interest: phenotypic homogeneity, protein productivity, and protein quality. More in particular, for 'optimal heterologous protein production', the CHO cell may comprise phenotypic homogeneity and protein productivity, or phenotypic homogeneity, and protein quality, or protein productivity, and protein quality, or phenotypic homogeneity, protein productivity, and protein quality and good protein titre. Optimal heterologous protein productivity can also only happen when the cell (i.e. the CHO cell) is stable and has high cell survivability.

As used herein, the term 'phenotypic homogeneity' refers to a state when all the cells in a population exhibit the same phenotype under a certain condition.

The term 'protein productivity' as used herein refers to a measure of the amount of protein made per viable cell at a single titer point. It is calculated by dividing the titer (mg) by the viable cell density (VCD or cells/ml), and the final measurement is represented as the amount of protein per cell (mg/cell).

The term 'protein quality' refers to the posttranslational modification of the protein that determines the efficacy and function of the protein. The modifications generally include phosphorylation, glycosylation, ubiquitination, methylation, acetylation, protein folding etc. For example, protein glycosylation is a critical quality attribute that modulates the efficacy, stability, and half-life of a therapeutic protein. Protein quality can be determined using Immunoprecipitation based techniques, Biochemical Assays, Mass spectrometry (MS) and the like.

As used herein, the term 'cell survivability' refers to the capability of a cell to be viable and perform cell proliferation. Cell viability is a measure of the proportion of live cells within a population. Cell proliferation refers to an increase in cell number due to cell division. The assays that are commonly used to test cell survivability include BrdU Cell Proliferation Assay, MTT Cell Proliferation Assays, trypan blue cell counting, and ATP Cell Viability Assays.

As used herein, the term 'cell exhaustion' refers to the state of the cell where it loses its capability to perform metabolic activity including heterologous protein production. Cell exhaustion can be determined by Metabolite Detection Assays.

The term "significantly similar" as used herein, and in particular in context with the comparison of methylation profiles (such as the comparison between test profiles (from test subject(s) and reference profiles) shall mean a similarity observed by statistical means (i.e. by using bioinformatics) and/or also by observation using the eye. A significant similarity is observed for example if a test profile overlaps with a reference profile that is defined by multiple training samples

through multivariate statistical methods, such as Principal Component analysis or Multi-Dimensional Scaling. In particular, a test profile is significantly similar to the pre-determined reference profile if more than 50, 55, 60, 65, 70, 75, 80, 85, 90, 95 % of the methylation pattern/profile overlaps with that of the reference profile. A similarity of a test profile to more than one, such as two, three or even all reference profiles reduce the significance of the similarity.

The term "pre-determined reference profile" as used herein refers to a typical or standard methylation profile of the genomic material of a CHO cell line with a specific feature dependent on the context where the term is used. In one example, for a method of determining a CHO cell line that displays at least one phenotypic parameter according to any aspect of the present invention conferring the potential of optimal heterologous protein production on the cell line, the term "pre-determined reference profile" refers to a typical or standard methylation profile of the genomic material of the CHO cell line displaying one or more of the phenotypic parameters selected from the group consisting of optimal glucose consumption, optimal growth rate, optimal lactic acid production, and optimal ammonia accumulation. The pre-determined reference profile may be obtained from one or more reference CHO cell lines each expressing one or more phenotypic parameter.

The method according to this aspect of the present invention may use a methylation profile for a CHO cell line that has the potential for optimal heterologous protein production as the cell line may exhibit cell survivability, fitness, low cell exhaustion and good metabolic readouts. In particular, the method according to this aspect of the present invention provides a prognostic methylation profile for ideal parental cell lines prior to transgene introduction. This methylation profile may be used as a control.

The cell according to any aspect of the present invention may be obtained from a biological sample selected from the group consisting of saliva, blood, brain, sperm and any other tissue or sample that provides genomic DNA.

The virus according to any aspect of the present invention may be from the *Herpesviridae* and/or *Polyomaviridae* genus. In particular, the virus from the *Herpesviridae* genus is selected from the group consisting of *Simplexvirus*, *Mardivirus*, *Varicellovirus*, *Lymphocryptovirus*, *Cytomegalovirus*, *Muromegalovirus*, *Roseolovirus*, *Lymphocryptovirus* and *Rhadinovirus* and the virus from the *Polyomaviridae* genus is selected from the group consisting of *Alphapolyomavirus*, *Betapolyomavirus*, *Deltapolyomavirus*, *Epsilonpolyomavirus*, *Gammapolyomavirus* and *Zetapolyomavirus*. The virus from the *Herpesviridae* genus is selected from the group consisting of *Gallid alphaherpesvirus 2*, *Human Herpesvirus 1*, *Human Herpesvirus 2*, *Bovine Herpesvirus 1*, *Bovine Herpesvirus 5*, *Human Herpesvirus 3*, *Suid Herpesvirus 1*, *Human Herpesvirus 5*, *Murid Herpesvirus 1*, *Human Herpesvirus 6A*, *Human Herpesvirus 6B*, *Human Herpesvirus 7*, *Human Herpesvirus 4*, *Kaposi's sarcoma-associated HV*, and *Murid Herpesvirus 4*; and the virus from the *Polyomavirus* genus is selected from the group consisting of *Human polyomavirus 5*, *Human polyomavirus 8*, *Human polyomavirus 9*, *Human polyomavirus 12*, *Human polyomavirus 13*,

Human polyomavirus 1, Human polyomavirus 2, Human polyomavirus 3, Human polyomavirus 4, Human polyomavirus 6, Human polyomavirus 7, Human polyomavirus 10, Human polyomavirus 11, and Human polyomavirus 14. In particular, the virus may be *Human Herpesvirus 5* (also known as human cytomegalovirus), *Murid Herpesvirus 1* (also known as murine cytomegalovirus), or
5 *Simian virus 40*.

According to a further aspect of the present invention, there is provided a use of DNA-based array to determine:

- (a) the presence of integrated viral DNA in a mammalian cell; and
 - 10 (b) methylation status of at least one CpG site of the virus,
- thereby determining the status of the virus.

The mammalian cell may be an immortalised cell line derived from the mammal.

According to yet another aspect of the present invention, there is provided a DNA-based array for determining

- 15 (a) the presence of integrated viral DNA in a mammalian cell; and
 - (b) methylation status of at least one CpG site of the virus,
- thereby determining the status of the virus.

EXAMPLES

- 20 The foregoing describes preferred embodiments, which, as will be understood by those skilled in the art, may be subject to variations or modifications in design, construction or operation without departing from the scope of the claims. These variations, for instance, are intended to be covered by the scope of the claims.

25 Example 1

Detection of viral promoter in CHO

- CHO cells have Cytomegalovirus Major Immediate-Early and enhancer (CMV) promoter and simian vacuolating virus 40 (SV40) promoters which are the most frequently used promoters for the production of recombinant proteins in CHO cells. CHO DG44 (Thermo Fisher) is a parental CHO
30 cell line which does not contact any viral promoter integrated into the genome. Humira431 (A*STAR Bioprocessing Technology Institute) cell line is a transgenic cell line derived from CHO DG44 cells which express biosimilar Humira from the modified CMV promoter.

DNA Extraction

- 35 DNA was extracted using the PureLink Genomic DNA Isolation Minikit kit (Invitrogen), including RNAase treatment following the manufacturer's instructions. DNA quantity was measured by PicoGreen assay and DNA quality is assessed via NanoDrop (Thermo Scientific) to ensure the A260/280 ratio is ≤ 1.8 . A small amount of sample was then also analysed using automated electrophoresis on TapeStation (Agilent) to ensure each sample contains high molecular weight
40 DNA.

Bisulfite Conversion and BeadChip Analysis

The genomic DNA samples were then subjected to bisulfite conversion using the EZ DNA Methylation-Gold™ Kit (Zymo Research). The methylation levels were then quantified using our customized methylation BeadChip kits (Illumina) which can analyze over 50,000 methylation sites quantitatively across the genome at single-nucleotide resolution.

5

Data processing:

The customized chip array data processing was performed in R version 4.1.2 using sesame version 1.14.2. DNA methylation level for each site was calculated as methylation β -value. Beta values were defined as methylated signal/(methylated signal + unmethylated signal). It

10 can be computed using getBetas function. The SeSAmE pipeline (Zhou et al. 2018) was used to generate normalized β -values and for quality control. Low intensity- based detection calling and making (based on p-value) was done with pOOBAH. Background subtraction based on normal-exponential deconvolution using out-of-band probes noob (Triche et al. 2013) and optionally with extra bleed-through subtraction were also implemented.

15

Percentage of viral probes which pass the p-value form the data indicated the presence and type of viral integrated in the CHO genome.

Results

20 3 CHO DG44 samples and 3 CHO Humira431 samples were run on TALOS. CHO Humira431 cell line contains CMV vector while CHO DG44 does not. The result shows that only CHO Humira431 cells can be detected by the probes targeted to CMV. In total, there are 538 CMV probes on TALOS1. Table 1 lists the number of viral probes that can be successfully detected.

Table 1: Number of successfully detected CMV probes in 3 CHO DG44 samples and 3 CHO

25 Humira431 samples

samples	Successfully detected probes (p value <0.05)
DG44_sample1	0
DG44_sample2	0
DG44_sample3	0
Humira_sample1	538
Humira_sample2	537
Humira_sample3	538

Example 2

Quantifying the methylation levels of viral promoters in CHO genomes

30 The protein production variability in CHO cells has been associated with DNA methylation mediated regulation of Cytomegalovirus Major Immediate-Early and enhancer (CMV) promoter and simian vacuolating virus 40 (SV40) promoters which are the most frequently used promoters for the production of recombinant proteins in CHO cells. An increase in DNA methylation on the CpG sites of the viral promoter results in transgene silencing at transcription levels

DNA extraction, bisulphite conversion, BeadChip analysis, quality control, data processing and differential methylation analysis were carried out as outlined in Example 1.

- 5 Analysing the level of DNA methylation on the CpG sites on the viral promoter can predict the expression of a transgene in CHO cells.

3 CHO DG44 samples and 3 CHO Humira431 samples were run on TALOS. CHO Humira431 cell line contains CMV vector while CHO DG44 does not. Table 2 lists all the probes specific to CMV
10 that can be used for CMV detection together with the beta values in CHO Humira431 samples.

Table 2: Beta value of CMV probes in 3 CHO Humira431 samples

Specific Probes	Humira_sample1	Humira_sample2	Humira_sample3				
cg106CG_TC1_R_CMV1	0.015	0.016	0.016	ch145CA_BC2_F_CMV1	0.097	0.106	0.112
cg125CG_TC1_F_CMV1	0.015	0.016	0.017	ch162CC_TC2_F_CMV1	0.030	0.029	0.026
cg132CG_BC1_F_CMV1	0.015	0.016	0.016	ch164CA_BC2_F_CMV1	0.110	0.121	0.127
cg142CG_BC1_F_CMV1	0.021	0.019	0.017	ch174CC_BC2_F_CMV1	0.045	0.040	0.043
cg142CG_BC2_F_CMV1	0.067	0.054	0.047	ch175CA_BC2_F_CMV1	0.072	0.077	0.084
cg154CG_BC2_F_CMV1	0.120	0.105	0.111	ch181GG_BC1_F_CMV1	0.958	0.951	0.959
cg172CG_TC1_F_CMV1	0.031	0.033	0.028	ch184CT_TC2_R_CMV1	0.915	0.904	0.920
cg172CG_TC2_F_CMV1	0.082	0.084	0.083	ch189CA_TC1_R_CMV1	0.869	0.869	0.880
cg195CG_BC2_R_CMV1	0.117	0.109	0.112	ch189CA_TC2_R_CMV1	0.948	0.954	0.947
cg195CG_TC1_F_CMV1	0.052	0.052	0.050	ch198CA_TC2_R_CMV1	0.937	0.932	0.929
cg195CG_TC2_F_CMV1	0.125	0.121	0.114	ch203GG_BC2_F_CMV1	0.816	0.823	0.815
cg216CG_BC1_R_CMV1	0.028	0.029	0.027	ch205TG_BC1_F_CMV1	0.031	0.039	0.034
cg216CG_TC1_F_CMV1	0.157	0.155	0.143	ch205TG_BC2_F_CMV1	0.152	0.168	0.165
cg216CG_TC2_F_CMV1	0.093	0.091	0.087	ch205TG_TC1_R_CMV1	0.069	0.076	0.073
cg265CG_TC1_F_CMV1	0.016	0.017	0.016	ch205TG_TC2_R_CMV1	0.024	0.025	0.027
cg265CG_TC2_F_CMV1	0.048	0.066	0.046	ch206GG_BC2_F_CMV1	0.851	0.851	0.852
cg278CG_BC1_R_CMV1	0.039	0.039	0.031	ch206GG_TC2_R_CMV1	0.096	0.097	0.086
cg278CG_BC2_R_CMV1	0.090	0.087	0.092	ch208AG_BC1_F_CMV1	0.032	0.037	0.031
cg278CG_TC1_F_CMV1	0.026	0.022	0.026	ch217GG_BC1_F_CMV1	0.944	0.947	0.943
cg278CG_TC2_F_CMV1	0.101	0.105	0.103	ch217GG_BC2_F_CMV1	0.940	0.944	0.944
cg287CG_BC1_R_CMV1	0.046	0.052	0.051	ch217GG_TC1_R_CMV1	0.024	0.027	0.033
cg287CG_BC2_R_CMV1	0.115	0.106	0.089	ch217GG_TC2_R_CMV1	0.075	0.084	0.078
cg287CG_TC2_F_CMV1	0.154	0.166	0.164	ch223CT_TC2_F_CMV1	0.041	0.032	0.030
cg299CG_BC1_F_CMV1	0.145	0.139	0.117	ch226CC_BC1_R_CMV1	0.848	0.854	0.826
cg299CG_BC2_F_CMV1	0.095	0.102	0.116	ch226CC_TC2_F_CMV1	0.041	0.038	0.045
cg371CG_BC1_R_CMV1	0.069	0.073	0.095	ch228CA_BC2_F_CMV1	0.104	0.105	0.083
cg371CG_BC2_R_CMV1	0.134	0.134	0.131	ch228CA_TC2_R_CMV1	0.907	0.896	0.908
cg414CG_TC1_R_CMV1	0.068	0.072	0.073	ch230CT_BC2_F_CMV1	0.049	0.051	0.055
cg435CG_BC1_R_CMV1	0.025	0.024	0.018	ch230CT_TC1_R_CMV1	0.906	0.884	0.899
cg464CG_BC2_F_CMV1	0.122	0.137	0.134	ch230CT_TC2_R_CMV1	0.916	0.919	0.919
cg464CG_TC2_R_CMV1	0.091	0.091	0.096	ch235CA_TC2_R_CMV1	0.945	0.946	0.953
cg498CG_BC1_R_CMV1	0.134	0.119	0.103	ch236AG_BC1_F_CMV1	0.024	0.028	0.023
cg498CG_BC2_R_CMV1	0.108	0.102	0.086	ch236AG_TC1_R_CMV1	0.126	0.121	0.116
cg498CG_BC2_F_CMV1	0.107	0.107	0.107	ch236AG_TC2_R_CMV1	0.088	0.093	0.088
cg516CG_TC2_R_CMV1	0.097	0.103	0.089	ch245AG_BC1_F_CMV1	0.017	0.017	0.015
cg527CG_BC1_R_CMV1	0.023	0.017	0.018	ch247TG_BC1_F_CMV1	0.010	0.014	0.010
cg78CG_BC1_R_CMV1	0.029	0.025	0.023	ch247TG_BC2_F_CMV1	0.056	0.076	0.061
cg78CG_BC2_R_CMV1	0.081	0.083	0.071	ch261AG_BC1_F_CMV1	0.013	0.014	0.014
cg80CG_TC1_R_CMV1	0.016	0.017	0.019	ch267CC_TC2_F_CMV1	0.040	0.054	0.044
cg80CG_TC2_R_CMV1	0.050	0.043	0.045	ch268CC_BC2_F_CMV1	0.062	0.066	0.055
cg94CG_BC1_R_CMV1	0.042	0.055	0.038	ch269CC_TC2_R_CMV1	0.840	0.851	0.866
cg94CG_BC2_R_CMV1	0.098	0.091	0.083	ch270CC_BC2_F_CMV1	0.027	0.033	0.038
ch104CC_BC2_R_CMV1	0.953	0.955	0.954	ch271CT_BC2_F_CMV1	0.042	0.045	0.047
ch114TG_BC1_F_CMV1	0.018	0.018	0.017	ch271CT_TC2_R_CMV1	0.892	0.885	0.881
				ch281CA_BC2_F_CMV1	0.161	0.164	0.163
				ch288GG_BC1_F_CMV1	0.859	0.874	0.860

ch288GG_BC2_F_CMV1	0.859	0.845	0.858	ch526CC_BC1_R_CMV1	0.878	0.897	0.896
ch288GG_TC1_R_CMV1	0.076	0.078	0.082	ch529CC_TC2_R_CMV1	0.886	0.911	0.913
ch288GG_TC2_R_CMV1	0.115	0.102	0.101	ch64CA_TC2_R_CMV1	0.905	0.928	0.911
ch297CC_BC2_R_CMV1	0.922	0.923	0.935	ch73AG_TC2_R_CMV1	0.020	0.022	0.021
ch298CC_TC2_F_CMV1	0.058	0.059	0.055	ch90CT_BC2_R_CMV1	0.959	0.955	0.956
ch301CC_BC2_F_CMV1	0.095	0.096	0.096	ch95GG_TC1_R_CMV1	0.046	0.048	0.054
ch302CT_BC2_F_CMV1	0.101	0.099	0.088	ch95GG_TC2_R_CMV1	0.088	0.106	0.079
ch303TG_BC1_F_CMV1	0.309	0.290	0.295	rs100AT_BC2_R_CMV1	0.024	0.022	0.027
ch303TG_BC2_F_CMV1	0.185	0.183	0.179	rs101TG_BC1_R_CMV1	0.017	0.018	0.019
ch306CA_BC2_F_CMV1	0.145	0.144	0.132	rs101TG_BC2_R_CMV1	0.027	0.027	0.023
ch313CC_TC2_F_CMV1	0.050	0.048	0.046	rs102GG_BC1_R_CMV1	0.109	0.120	0.111
ch314CC_BC2_F_CMV1	0.051	0.063	0.060	rs102GG_BC2_R_CMV1	0.084	0.090	0.083
ch316AG_BC1_F_CMV1	0.158	0.158	0.170	rs103GC_BC1_R_CMV1	0.048	0.047	0.078
ch316AG_BC2_F_CMV1	0.073	0.065	0.072	rs107GC_TC2_R_CMV1	0.095	0.093	0.085
ch320CA_BC2_R_CMV1	0.905	0.911	0.855	rs111GG_BC1_R_CMV1	0.116	0.137	0.127
ch325CC_TC2_R_CMV1	0.930	0.928	0.926	rs115GA_BC1_R_CMV1	0.091	0.066	0.077
ch343CT_TC2_F_CMV1	0.033	0.035	0.033	rs116AC_BC1_R_CMV1	0.116	0.098	0.079
ch349AG_BC1_F_CMV1	0.041	0.043	0.042	rs123AA_BC1_F_CMV1	0.016	0.016	0.014
ch368CA_BC1_R_CMV1	0.880	0.886	0.882	rs124AC_TC1_F_CMV1	0.016	0.013	0.014
ch373CT_TC2_R_CMV1	0.906	0.904	0.925	rs137AT_BC1_F_CMV1	0.015	0.017	0.015
ch379CC_BC1_R_CMV1	0.888	0.880	0.886	rs138TT_TC2_F_CMV1	0.030	0.025	0.035
ch379CC_BC2_R_CMV1	0.930	0.938	0.939	rs139TG_TC1_F_CMV1	0.014	0.015	0.012
ch380CA_TC1_R_CMV1	0.020	0.022	0.021	rs141AC_TC1_F_CMV1	0.010	0.009	0.008
ch380CA_TC2_R_CMV1	0.073	0.069	0.073	rs141AC_TC2_F_CMV1	0.029	0.032	0.035
ch417GG_BC1_F_CMV1	0.610	0.627	0.683	rs144TC_BC1_F_CMV1	0.293	0.305	0.312
ch424GG_TC1_R_CMV1	0.033	0.033	0.029	rs144TC_BC2_F_CMV1	0.027	0.029	0.022
ch438GG_BC2_F_CMV1	0.927	0.906	0.924	rs146AA_BC2_F_CMV1	0.027	0.030	0.028
ch444CC_TC2_F_CMV1	0.030	0.028	0.025	rs147AT_BC1_F_CMV1	0.014	0.017	0.014
ch445CA_BC2_F_CMV1	0.093	0.089	0.084	rs147AT_BC2_F_CMV1	0.030	0.026	0.026
ch447AG_BC1_F_CMV1	0.056	0.064	0.082	rs148TA_BC1_F_CMV1	0.202	0.201	0.190
ch447AG_BC2_F_CMV1	0.051	0.057	0.061	rs149AA_TC1_F_CMV1	0.016	0.017	0.015
ch447AG_TC2_R_CMV1	0.026	0.026	0.023	rs150AT_TC2_F_CMV1	0.041	0.053	0.049
ch450CT_BC2_R_CMV1	0.903	0.926	0.919	rs151TG_TC2_F_CMV1	0.078	0.080	0.083
ch452CC_TC2_F_CMV1	0.022	0.029	0.031	rs153AC_TC1_F_CMV1	0.010	0.012	0.011
ch455CC_TC2_F_CMV1	0.025	0.027	0.025	rs155GT_BC1_F_CMV1	0.619	0.617	0.634
ch456CC_BC2_F_CMV1	0.032	0.031	0.034	rs155GT_BC2_F_CMV1	0.912	0.941	0.926
ch457CC_BC2_F_CMV1	0.032	0.030	0.034	rs156TA_BC1_F_CMV1	0.357	0.376	0.373
ch474AG_TC2_R_CMV1	0.038	0.052	0.048	rs156TA_BC2_F_CMV1	0.064	0.076	0.065
ch478TG_TC1_R_CMV1	0.033	0.034	0.032	rs157AT_TC1_F_CMV1	0.026	0.026	0.031
ch486CA_TC2_R_CMV1	0.945	0.951	0.953	rs157AT_TC2_F_CMV1	0.023	0.025	0.022
ch488CC_TC2_R_CMV1	0.953	0.954	0.956	rs158TG_TC1_F_CMV1	0.018	0.017	0.018
ch489CA_TC2_R_CMV1	0.929	0.917	0.920	rs158TG_TC2_F_CMV1	0.060	0.059	0.045
ch495CA_BC1_R_CMV1	0.905	0.908	0.914	rs159GT_TC2_F_CMV1	0.954	0.956	0.964
ch495CA_BC2_R_CMV1	0.916	0.915	0.919	rs160TT_TC2_F_CMV1	0.029	0.029	0.032
ch503CT_TC2_R_CMV1	0.874	0.864	0.864	rs165AT_BC1_F_CMV1	0.077	0.072	0.075
ch521CA_BC2_R_CMV1	0.956	0.948	0.946	rs171AC_TC1_F_CMV1	0.014	0.016	0.014

rs173GC_BC2_F_CMV1	0.895	0.915	0.901	rs204GT_BC2_R_CMV1	0.040	0.040	0.051
rs176AA_BC2_F_CMV1	0.034	0.035	0.040	rs204GT_TC1_F_CMV1	0.861	0.883	0.877
rs177AT_BC2_F_CMV1	0.025	0.025	0.024	rs204GT_TC2_F_CMV1	0.837	0.837	0.831
rs179AG_BC1_R_CMV1	0.055	0.064	0.083	rs207GA_BC2_F_CMV1	0.879	0.872	0.876
rs179AG_TC2_F_CMV1	0.082	0.075	0.079	rs207GA_TC1_R_CMV1	0.365	0.363	0.369
rs182GA_BC1_F_CMV1	0.948	0.940	0.939	rs207GA_TC2_R_CMV1	0.021	0.023	0.024
rs182GA_BC2_F_CMV1	0.941	0.925	0.919	rs209GT_BC2_F_CMV1	0.910	0.909	0.909
rs182GA_TC1_R_CMV1	0.047	0.053	0.046	rs209GT_TC1_R_CMV1	0.403	0.374	0.365
rs183AC_BC1_F_CMV1	0.019	0.015	0.018	rs209GT_TC2_R_CMV1	0.019	0.021	0.020
rs183AC_TC1_R_CMV1	0.039	0.064	0.049	rs210TA_BC1_F_CMV1	0.388	0.387	0.383
rs185TT_BC1_F_CMV1	0.337	0.377	0.335	rs210TA_TC2_R_CMV1	0.025	0.023	0.023
rs185TT_BC2_F_CMV1	0.069	0.067	0.086	rs211AT_BC1_F_CMV1	0.069	0.053	0.056
rs185TT_TC2_R_CMV1	0.046	0.048	0.049	rs211AT_BC2_F_CMV1	0.039	0.039	0.039
rs186TT_BC1_R_CMV1	0.019	0.021	0.022	rs211AT_TC1_R_CMV1	0.380	0.391	0.394
rs186TT_TC1_F_CMV1	0.396	0.411	0.385	rs211AT_TC2_R_CMV1	0.036	0.031	0.036
rs187TC_BC1_R_CMV1	0.013	0.016	0.013	rs212TT_BC1_R_CMV1	0.030	0.023	0.023
rs187TC_TC1_F_CMV1	0.331	0.316	0.306	rs212TT_BC2_R_CMV1	0.023	0.022	0.018
rs187TC_TC2_F_CMV1	0.031	0.030	0.034	rs213TT_BC1_R_CMV1	0.021	0.025	0.017
rs190AT_BC1_F_CMV1	0.036	0.045	0.038	rs213TT_TC1_F_CMV1	0.330	0.308	0.313
rs190AT_BC2_F_CMV1	0.031	0.030	0.029	rs214TA_BC2_R_CMV1	0.019	0.017	0.020
rs190AT_TC2_R_CMV1	0.050	0.044	0.050	rs214TA_TC2_F_CMV1	0.136	0.132	0.123
rs191TT_TC1_F_CMV1	0.374	0.379	0.362	rs215AC_BC1_R_CMV1	0.403	0.404	0.382
rs191TT_TC2_F_CMV1	0.041	0.034	0.046	rs215AC_TC1_F_CMV1	0.065	0.063	0.059
rs192TG_BC2_R_CMV1	0.038	0.036	0.037	rs215AC_TC2_F_CMV1	0.053	0.056	0.052
rs192TG_TC1_F_CMV1	0.023	0.026	0.017	rs218GT_BC1_F_CMV1	0.961	0.971	0.967
rs192TG_TC2_F_CMV1	0.110	0.117	0.135	rs218GT_BC2_F_CMV1	0.958	0.960	0.968
rs193GA_BC1_R_CMV1	0.377	0.391	0.378	rs219TA_BC2_F_CMV1	0.056	0.057	0.062
rs193GA_BC2_R_CMV1	0.155	0.145	0.141	rs220AA_BC1_F_CMV1	0.014	0.018	0.016
rs193GA_TC2_F_CMV1	0.853	0.856	0.856	rs221AA_BC1_R_CMV1	0.022	0.022	0.020
rs194AC_BC1_R_CMV1	0.409	0.406	0.403	rs221AA_BC2_R_CMV1	0.027	0.039	0.026
rs194AC_BC2_R_CMV1	0.042	0.037	0.039	rs221AA_TC2_F_CMV1	0.039	0.034	0.039
rs196GT_BC1_F_CMV1	0.754	0.744	0.762	rs222AC_BC1_R_CMV1	0.179	0.182	0.193
rs196GT_BC2_F_CMV1	0.813	0.810	0.813	rs222AC_TC2_F_CMV1	0.039	0.037	0.040
rs196GT_TC1_R_CMV1	0.046	0.055	0.042	rs224TG_TC2_F_CMV1	0.132	0.135	0.136
rs196GT_TC2_R_CMV1	0.102	0.106	0.096	rs225GC_TC2_F_CMV1	0.846	0.827	0.841
rs197TC_BC1_F_CMV1	0.492	0.490	0.485	rs229AC_BC1_F_CMV1	0.070	0.079	0.068
rs197TC_TC1_R_CMV1	0.013	0.011	0.013	rs229AC_BC2_F_CMV1	0.057	0.051	0.061
rs197TC_TC2_R_CMV1	0.034	0.034	0.040	rs229AC_TC1_R_CMV1	0.024	0.022	0.023
rs199AA_TC1_R_CMV1	0.017	0.022	0.018	rs229AC_TC2_R_CMV1	0.117	0.111	0.109
rs199AA_TC2_R_CMV1	0.111	0.114	0.102	rs231TT_BC2_R_CMV1	0.086	0.075	0.092
rs200AT_BC1_R_CMV1	0.384	0.416	0.390	rs231TT_TC1_F_CMV1	0.411	0.431	0.393
rs200AT_TC2_F_CMV1	0.047	0.046	0.049	rs231TT_TC2_F_CMV1	0.074	0.071	0.076
rs201TG_BC1_R_CMV1	0.040	0.045	0.045	rs232TG_BC2_R_CMV1	0.043	0.047	0.049
rs201TG_TC2_F_CMV1	0.153	0.141	0.145	rs233GG_BC2_R_CMV1	0.140	0.144	0.118
rs202GG_BC1_R_CMV1	0.408	0.368	0.360	rs233GG_TC1_F_CMV1	0.702	0.702	0.711
rs204GT_BC1_R_CMV1	0.390	0.392	0.340	rs234GC_BC1_F_CMV1	0.843	0.851	0.860

rs234GC_BC2_F_CMV1	0.956	0.962	0.967	rs284TG_BC1_R_CMV1	0.038	0.042	0.038
rs234GC_TC1_R_CMV1	0.354	0.369	0.333	rs284TG_BC2_R_CMV1	0.031	0.028	0.027
rs237GT_BC2_F_CMV1	0.958	0.952	0.959	rs284TG_TC1_F_CMV1	0.018	0.016	0.017
rs237GT_TC1_R_CMV1	0.356	0.392	0.385	rs284TG_TC2_F_CMV1	0.091	0.095	0.105
rs237GT_TC2_R_CMV1	0.054	0.060	0.048	rs285GA_BC2_R_CMV1	0.151	0.149	0.153
rs238TA_BC1_F_CMV1	0.166	0.182	0.173	rs285GA_TC2_F_CMV1	0.809	0.806	0.815
rs238TA_TC2_R_CMV1	0.029	0.032	0.033	rs286AC_BC1_R_CMV1	0.376	0.386	0.375
rs239AC_BC1_F_CMV1	0.023	0.024	0.023	rs286AC_BC2_R_CMV1	0.032	0.038	0.041
rs239AC_TC1_R_CMV1	0.331	0.397	0.390	rs286AC_TC2_F_CMV1	0.100	0.109	0.111
rs239AC_TC2_R_CMV1	0.082	0.077	0.076	rs289GT_BC1_F_CMV1	0.929	0.937	0.944
rs241AT_BC1_F_CMV1	0.017	0.021	0.018	rs289GT_BC2_F_CMV1	0.908	0.917	0.933
rs242TC_BC1_F_CMV1	0.220	0.244	0.244	rs290TA_BC1_F_CMV1	0.449	0.459	0.449
rs242TC_BC2_F_CMV1	0.028	0.030	0.027	rs290TA_BC2_F_CMV1	0.101	0.086	0.106
rs246GT_TC2_F_CMV1	0.960	0.953	0.949	rs290TA_TC1_R_CMV1	0.017	0.017	0.017
rs249TA_BC1_F_CMV1	0.075	0.066	0.070	rs291AA_BC1_F_CMV1	0.089	0.089	0.088
rs250AT_TC1_F_CMV1	0.018	0.020	0.022	rs291AA_BC2_F_CMV1	0.046	0.047	0.043
rs250AT_TC2_F_CMV1	0.034	0.033	0.034	rs291AA_TC1_R_CMV1	0.164	0.169	0.166
rs251TC_BC1_F_CMV1	0.063	0.082	0.077	rs292AA_BC1_R_CMV1	0.166	0.189	0.170
rs253AT_BC1_F_CMV1	0.021	0.016	0.015	rs292AA_TC1_F_CMV1	0.070	0.061	0.072
rs255AT_TC1_F_CMV1	0.014	0.016	0.016	rs292AA_TC2_F_CMV1	0.046	0.038	0.044
rs272TA_BC1_F_CMV1	0.172	0.174	0.162	rs293AT_BC1_R_CMV1	0.078	0.087	0.082
rs272TA_BC2_F_CMV1	0.157	0.161	0.170	rs294TG_BC1_R_CMV1	0.024	0.025	0.022
rs273AT_TC1_R_CMV1	0.467	0.447	0.451	rs295GG_TC1_F_CMV1	0.870	0.849	0.871
rs273AT_TC2_R_CMV1	0.071	0.079	0.079	rs296GC_BC1_R_CMV1	0.041	0.045	0.043
rs274TT_BC1_R_CMV1	0.103	0.115	0.099	rs296GC_TC2_F_CMV1	0.827	0.819	0.835
rs274TT_TC1_F_CMV1	0.212	0.216	0.217	rs300GC_TC1_R_CMV1	0.016	0.017	0.017
rs274TT_TC2_F_CMV1	0.042	0.048	0.041	rs304GG_TC2_F_CMV1	0.871	0.875	0.868
rs275TG_TC1_F_CMV1	0.014	0.013	0.014	rs305GC_BC1_F_CMV1	0.914	0.921	0.929
rs275TG_TC2_F_CMV1	0.092	0.103	0.101	rs305GC_BC2_F_CMV1	0.915	0.914	0.922
rs276GA_BC1_R_CMV1	0.449	0.445	0.439	rs307AT_BC1_F_CMV1	0.134	0.169	0.142
rs276GA_BC2_R_CMV1	0.095	0.101	0.104	rs307AT_BC2_F_CMV1	0.063	0.059	0.068
rs276GA_TC1_F_CMV1	0.341	0.348	0.318	rs308TT_BC1_F_CMV1	0.469	0.473	0.488
rs276GA_TC2_F_CMV1	0.905	0.897	0.906	rs310AT_TC2_F_CMV1	0.061	0.064	0.060
rs277AC_TC1_F_CMV1	0.009	0.010	0.009	rs311TG_TC1_F_CMV1	0.059	0.061	0.058
rs279GT_BC2_F_CMV1	0.843	0.856	0.867	rs311TG_TC2_F_CMV1	0.111	0.102	0.093
rs279GT_TC1_R_CMV1	0.041	0.048	0.041	rs312GC_TC1_F_CMV1	0.842	0.845	0.865
rs279GT_TC2_R_CMV1	0.090	0.082	0.077	rs312GC_TC2_F_CMV1	0.892	0.895	0.907
rs280TC_BC1_F_CMV1	0.298	0.303	0.319	rs317GT_BC1_F_CMV1	0.788	0.787	0.795
rs280TC_TC1_R_CMV1	0.012	0.013	0.012	rs317GT_BC2_F_CMV1	0.937	0.943	0.938
rs280TC_TC2_R_CMV1	0.039	0.042	0.040	rs318TA_BC1_F_CMV1	0.503	0.493	0.477
rs282AA_BC1_F_CMV1	0.030	0.033	0.038	rs318TA_BC2_F_CMV1	0.110	0.106	0.106
rs282AA_BC2_F_CMV1	0.069	0.068	0.080	rs318TA_TC1_R_CMV1	0.022	0.024	0.034
rs282AA_TC1_R_CMV1	0.020	0.022	0.025	rs319AC_BC1_F_CMV1	0.159	0.155	0.160
rs282AA_TC2_R_CMV1	0.094	0.092	0.097	rs319AC_TC1_R_CMV1	0.252	0.253	0.233
rs283AT_BC1_R_CMV1	0.391	0.379	0.416	rs321AT_BC1_R_CMV1	0.020	0.017	0.021
rs283AT_TC1_F_CMV1	0.025	0.029	0.022	rs322TG_BC1_R_CMV1	0.027	0.033	0.032

rs323GA_BC1_R_CMV1	0.258	0.266	0.245	rs427TT_TC1_F_CMV1	0.306	0.291	0.302
rs323GA_BC2_R_CMV1	0.150	0.150	0.153	rs428TG_BC1_R_CMV1	0.020	0.022	0.023
rs323GA_TC1_F_CMV1	0.915	0.918	0.910	rs432TC_BC1_F_CMV1	0.326	0.336	0.314
rs323GA_TC2_F_CMV1	0.911	0.908	0.902	rs432TC_TC1_R_CMV1	0.011	0.013	0.013
rs324AC_BC1_F_CMV1	0.018	0.020	0.018	rs434AC_BC1_R_CMV1	0.012	0.014	0.012
rs324AC_TC2_R_CMV1	0.056	0.051	0.042	rs436GG_BC1_R_CMV1	0.028	0.025	0.018
rs327TT_BC1_F_CMV1	0.260	0.266	0.255	rs439GA_BC1_F_CMV1	0.954	0.963	0.954
rs328TA_TC1_F_CMV1	0.360	0.344	0.358	rs439GA_BC2_F_CMV1	0.904	0.911	0.917
rs334AC_BC1_F_CMV1	0.026	0.022	0.023	rs439GA_TC1_R_CMV1	0.045	0.041	0.033
rs337TT_TC1_F_CMV1	0.142	0.151	0.159	rs441TT_TC1_R_CMV1	0.021	0.017	0.020
rs341TA_BC2_F_CMV1	0.050	0.055	0.059	rs443TC_BC1_R_CMV1	0.015	0.014	0.014
rs342AC_BC2_F_CMV1	0.033	0.028	0.035	rs446AA_BC1_F_CMV1	0.078	0.090	0.085
rs345TG_TC1_F_CMV1	0.022	0.021	0.019	rs446AA_TC1_R_CMV1	0.018	0.018	0.016
rs345TG_TC2_F_CMV1	0.061	0.074	0.067	rs448GT_TC2_F_CMV1	0.919	0.916	0.918
rs363TT_BC1_R_CMV1	0.027	0.036	0.025	rs449TC_BC1_F_CMV1	0.383	0.363	0.389
rs364TA_BC1_R_CMV1	0.046	0.055	0.040	rs451TC_BC1_R_CMV1	0.018	0.017	0.012
rs364TA_BC2_R_CMV1	0.046	0.037	0.040	rs451TC_TC2_F_CMV1	0.023	0.026	0.030
rs365AG_BC1_R_CMV1	0.406	0.408	0.409	rs454AC_TC1_F_CMV1	0.122	0.110	0.113
rs365AG_BC2_R_CMV1	0.097	0.089	0.088	rs454AC_TC2_F_CMV1	0.044	0.042	0.049
rs366GT_BC1_R_CMV1	0.388	0.362	0.353	rs459AT_BC1_F_CMV1	0.084	0.090	0.096
rs366GT_BC2_R_CMV1	0.030	0.030	0.033	rs459AT_BC2_F_CMV1	0.037	0.034	0.028
rs367TC_TC1_R_CMV1	0.019	0.020	0.018	rs460TT_TC1_F_CMV1	0.435	0.432	0.425
rs367TC_TC2_R_CMV1	0.042	0.039	0.041	rs460TT_TC2_F_CMV1	0.032	0.030	0.044
rs369AT_BC1_R_CMV1	0.041	0.051	0.035	rs461TG_BC1_R_CMV1	0.114	0.130	0.105
rs369AT_BC2_R_CMV1	0.072	0.078	0.062	rs461TG_TC1_F_CMV1	0.026	0.034	0.025
rs370TC_BC2_R_CMV1	0.048	0.044	0.041	rs461TG_TC2_F_CMV1	0.124	0.118	0.121
rs372GC_TC1_R_CMV1	0.068	0.073	0.073	rs462GA_BC1_R_CMV1	0.456	0.469	0.466
rs372GC_TC2_R_CMV1	0.130	0.114	0.123	rs462GA_TC2_F_CMV1	0.864	0.857	0.863
rs374TA_TC1_R_CMV1	0.020	0.026	0.023	rs463AC_BC1_R_CMV1	0.477	0.492	0.462
rs374TA_TC2_R_CMV1	0.111	0.110	0.098	rs463AC_TC1_F_CMV1	0.017	0.019	0.015
rs375AT_TC1_R_CMV1	0.312	0.333	0.325	rs465GT_BC2_F_CMV1	0.878	0.879	0.881
rs375AT_TC2_R_CMV1	0.120	0.116	0.113	rs465GT_TC2_R_CMV1	0.092	0.085	0.097
rs377TA_BC1_R_CMV1	0.029	0.033	0.028	rs466TC_BC1_F_CMV1	0.456	0.472	0.466
rs377TA_BC2_R_CMV1	0.053	0.045	0.043	rs466TC_BC2_F_CMV1	0.052	0.061	0.055
rs378AC_BC1_R_CMV1	0.412	0.395	0.425	rs466TC_TC1_R_CMV1	0.011	0.012	0.011
rs378AC_BC2_R_CMV1	0.084	0.078	0.080	rs470TG_BC1_R_CMV1	0.051	0.045	0.055
rs398GC_BC1_F_CMV1	0.686	0.676	0.686	rs473GA_TC1_R_CMV1	0.471	0.451	0.472
rs401GT_BC1_F_CMV1	0.749	0.752	0.750	rs473GA_TC2_R_CMV1	0.106	0.097	0.100
rs402TA_BC1_F_CMV1	0.394	0.398	0.413	rs475GT_BC1_F_CMV1	0.842	0.848	0.843
rs415GT_TC1_R_CMV1	0.060	0.063	0.062	rs475GT_BC2_F_CMV1	0.811	0.805	0.812
rs418GA_BC1_F_CMV1	0.690	0.729	0.704	rs475GT_TC1_R_CMV1	0.428	0.444	0.442
rs418GA_TC1_R_CMV1	0.118		0.138	rs475GT_TC2_R_CMV1	0.022	0.020	0.026
rs419AT_TC1_R_CMV1	0.150	0.146	0.124	rs476TT_BC1_F_CMV1	0.420	0.424	0.436
rs425GT_BC2_F_CMV1	0.905	0.886	0.889	rs476TT_TC1_R_CMV1	0.076	0.070	0.072
rs426TT_BC1_F_CMV1	0.289	0.326	0.292	rs477TT_BC1_R_CMV1	0.039	0.040	0.038
rs426TT_BC2_F_CMV1	0.055	0.048	0.050	rs477TT_BC2_R_CMV1	0.021	0.020	0.022

rs477TT_TC1_F_CMV1	0.368	0.379	0.384	rs65AT_BC1_F_CMV1	0.021	0.025	0.024
rs479GT_BC1_R_CMV1	0.446	0.432	0.456	rs65AT_TC1_R_CMV1	0.097	0.099	0.081
rs479GT_BC2_R_CMV1	0.020	0.022	0.022	rs65AT_TC2_R_CMV1	0.061	0.071	0.062
rs480TT_BC1_F_CMV1	0.213	0.216	0.278	rs66TA_TC1_R_CMV1	0.071	0.080	0.073
rs480TT_TC2_R_CMV1	0.017	0.020	0.020	rs67AT_TC1_R_CMV1	0.453	0.477	0.477
rs481TT_BC2_R_CMV1	0.017	0.020	0.020	rs67AT_TC2_R_CMV1	0.070	0.080	0.067
rs482TT_BC1_R_CMV1	0.045	0.053	0.045	rs68TA_BC1_R_CMV1	0.107	0.121	0.112
rs483TG_BC1_R_CMV1	0.043	0.041	0.038	rs68TA_BC2_R_CMV1	0.019	0.023	0.022
rs483TG_BC2_R_CMV1	0.023	0.025	0.023	rs70TG_BC1_R_CMV1	0.153	0.163	0.167
rs484GG_BC1_R_CMV1	0.455	0.460	0.473	rs75TT_BC2_R_CMV1	0.022	0.022	0.025
rs484GG_BC2_R_CMV1	0.053	0.054	0.057	rs76TC_BC1_R_CMV1	0.014	0.015	0.016
rs485GC_TC1_R_CMV1	0.455	0.483	0.468	rs79GC_TC1_R_CMV1	0.022	0.022	0.021
rs485GC_TC2_R_CMV1	0.026	0.029	0.027	rs79GC_TC2_R_CMV1	0.071	0.076	0.071
rs487AC_TC1_R_CMV1	0.029	0.025	0.024	rs81GT_TC1_R_CMV1	0.016	0.022	0.015
rs490AA_TC1_R_CMV1	0.298	0.294	0.280	rs81GT_TC2_R_CMV1	0.054	0.056	0.051
rs490AA_TC2_R_CMV1	0.116	0.118	0.119	rs83TA_BC1_R_CMV1	0.013	0.015	0.014
rs491AA_BC1_F_CMV1	0.015	0.015	0.013	rs84AC_TC1_R_CMV1	0.118	0.132	0.119
rs491AA_TC1_R_CMV1	0.418	0.433	0.409	rs84AC_TC2_R_CMV1	0.063	0.051	0.054
rs491AA_TC2_R_CMV1	0.047	0.054	0.050	rs86AT_TC1_R_CMV1	0.016	0.020	0.016
rs492AA_BC1_R_CMV1	0.420	0.440	0.455	rs87TA_TC1_R_CMV1	0.026	0.020	0.024
rs492AA_BC2_R_CMV1	0.030	0.035	0.032	rs87TA_TC2_R_CMV1	0.030	0.036	0.036
rs493AT_BC1_R_CMV1	0.430	0.413	0.451	rs88AA_BC1_R_CMV1	0.268	0.282	0.285
rs493AT_BC2_R_CMV1	0.036	0.037	0.042	rs88AA_BC2_R_CMV1	0.084	0.092	0.087
rs493AT_TC1_F_CMV1	0.036	0.022	0.028	rs89AC_BC2_R_CMV1	0.024	0.028	0.027
rs494TC_BC1_R_CMV1	0.023	0.021	0.022	rs89AC_TC1_F_CMV1	0.015	0.015	0.015
rs494TC_BC2_R_CMV1	0.027	0.030	0.032	rs91TT_TC1_R_CMV1	0.035	0.030	0.031
rs494TC_TC1_F_CMV1	0.088	0.102	0.098	rs91TT_TC2_R_CMV1	0.036	0.036	0.033
rs496AA_BC1_R_CMV1	0.067	0.047	0.075	rs92TA_BC2_R_CMV1	0.031	0.027	0.033
rs496AA_BC2_R_CMV1	0.066	0.060	0.071	rs93AC_BC1_R_CMV1	0.410	0.406	0.410
rs497AC_BC2_R_CMV1	0.047	0.053	0.051	rs93AC_BC2_R_CMV1	0.041	0.044	0.044
rs497AC_TC1_F_CMV1	0.013	0.013	0.012	rs97TA_TC2_R_CMV1	0.021	0.027	0.023
rs497AC_TC2_F_CMV1	0.025	0.028	0.024	rs98AA_TC1_R_CMV1	0.275	0.286	0.262
rs499GG_BC1_R_CMV1	0.134	0.131	0.127	rs99AA_BC1_R_CMV1	0.163	0.161	0.175
rs499GG_BC2_R_CMV1	0.102	0.092	0.096	rs99AA_BC2_R_CMV1	0.023	0.027	0.022
rs502AC_BC1_F_CMV1	0.015	0.014	0.015				
rs502AC_TC1_R_CMV1	0.230	0.249	0.225				
rs504TT_BC1_F_CMV1	0.043	0.060	0.048				
rs505TT_TC1_F_CMV1	0.067	0.057	0.056				
rs517GT_TC2_R_CMV1	0.093	0.095	0.105				
rs518TA_TC1_R_CMV1	0.017	0.020	0.022				
rs519AA_BC1_R_CMV1	0.205	0.193	0.218				
rs522AA_BC1_R_CMV1	0.026	0.024	0.022				
rs523AC_BC1_R_CMV1	0.177	0.187	0.179				
rs523AC_BC2_R_CMV1	0.060	0.059	0.069				
rs525TC_BC1_R_CMV1	0.024	0.027	0.024				
rs528GC_BC2_R_CMV1	0.113	0.119	0.132				

CLAIMS

1. A method of identifying integrated viral DNA in the genome of at least one mammalian test cell, the method comprising the steps of:
 - 5 (a) determining presence or absence of viral DNA in a DNA sample obtained from the mammalian test cell; and
 - (b) determining the methylation status of at least one CpG site in the viral DNA of the DNA sample obtained from the mammalian test cell,wherein the presence of the viral DNA after step (a) and hypermethylation in step (b) is
10 indicative of the viral DNA being integrated in the mammalian test cell and the virus being inactive respectively;
the presence of the viral DNA after step (a) and baseline methylation or hypomethylation in step (b) is indicative of the viral DNA being integrated in the mammalian test cell and the virus being active respectively; and
15 the absence of the viral DNA after step (a) is indicative of the viral DNA not being integrated in the mammalian test cell respectively; and
wherein the method is carried out on an array.
2. The method according to claim 1, wherein presence of specific CpG and non-CpG sites in
20 step (a) is indicative of the viral strain or variant thereof.
3. A method of identifying integrated viral DNA in the genome of at least one mammalian test cell, the method comprising the steps of:
 - 25 (a) determining presence or absence of one or more pre-selected CpG and/or non-CpG sites of the viral DNA in a DNA sample obtained from the mammalian test cell; and/or
 - (b) determining a test methylation profile of the pre-selected CpG and/or non-CpG sites within the DNA sample obtained from the mammalian test cell; andwherein the presence of one or more pre-selected CpG and/or non-CpG sites after step (a) and hypermethylation in (b) is indicative of the viral DNA being integrated in the mammalian
30 test cell but the virus being inactive respectively;
the presence of one or more pre-selected CpG and/or non-CpG sites after step (a) and baseline methylation or hypomethylation is indicative of the viral DNA being integrated in the mammalian test cell and the virus being active respectively; and
the absence of the one or more pre-selected CpG and/or non-CpG sites after step (a) is
35 indicative of the viral DNA not being integrated in the mammalian test cell respectively; and
wherein the method is carried out on a DNA- based array.
4. The method according to any one of the preceding claims, wherein the mammal is a mouse,
40 a rat, a guinea pig, a dog, a mini-pig, a human being, a cow, a sheep, a pig, a goat, a horse, a donkey, a mule, and a hamster.

5. The method according to any one of the preceding claims, wherein the mammalian cell is an immortalised cell line derived from the mammal.
6. The method according to claim 5, wherein the immortalised cell line is selected from the group consisting of CHO, BHK, Vero, HEK293, HEK 293T, HeLa cell, NS0 cell, Sp2/0 cell, and derivatives thereof.
7. The method according to any one of the preceding claims, wherein the mammalian cell is a CHO cell line.
8. The method according to claim 7, wherein there is a further step (c):
- (c) comparing the methylation status of the CpG site in the viral DNA from (b) with that of
- (i) a first CHO control cell where the viral DNA is integrated in the genome of the control and the first CHO control cell has optimal heterologous protein production;
- (ii) a second CHO control cell where the viral DNA is integrated in the genome of the control cell and the second CHO control cell has no optimal heterologous protein production;
- wherein significant similarity of the test methylation status of the CpG site with the first CHO control cell is indicative of the test CHO cell having optimal heterologous protein production, or with the second CHO control cell is indicative of the test CHO cell not having optimal heterologous protein production,
- wherein difference of the test methylation status of the CpG site with the first CHO control cell is indicative of the test CHO cell not having optimal heterologous protein production, or with the second or CHO control cell is indicative of the test CHO cell having optimal heterologous protein production.
9. The method according to any one of the preceding claims, wherein the integrated viral DNA is from at least one virus from the *Herpesviridae* and/or *Polyomaviridae* genus.
10. The method according to claim 9, wherein the virus from the *Herpesviridae* genus is selected from the group consisting of *Simplexvirus*, *Mardivirus*, *Varicellovirus*, *Lymphocryptovirus*, *Cytomegalovirus*, *Muromegalovirus*, *Roseolovirus*, *Lymphocryptovirus* and *Rhadinovirus* and the virus from the *Polyomaviridae* genus is selected from the group consisting of *Alphapolyomavirus*, *Betapolyomavirus*, *Deltapolyomavirus*, *Epsilonpolyomavirus*, *Gammapolyomavirus* and *Zetapolyomavirus*.
11. The method according to claim 9 or 10, wherein
- the virus from the *Herpesviridae* genus is selected from the group consisting of *Gallid alphaherpesvirus 2*, *Human Herpesvirus 1*, *Human Herpesvirus 2*, *Bovine Herpesvirus 1*, *Bovine Herpesvirus 5*, *Human Herpesvirus 3*, *Suid Herpesvirus 1*, *Human Herpesvirus 5*, *Murid Herpesvirus 1*, *Human Herpesvirus 6A*, *Human Herpesvirus 6B*, *Human*

- Herpesvirus 7, Human Herpesvirus 4, Kaposi's sarcoma-associated HV, and Murid Herpesvirus 4; and*
- the virus from the *Polyomavirus* genus is selected from the group consisting of *Human polyomavirus 5, Human polyomavirus 8, Human polyomavirus 9, Human polyomavirus 12, Human polyomavirus 13, Human polyomavirus 1, Human polyomavirus 2, Human polyomavirus 3, Human polyomavirus 4, Human polyomavirus 6, Human polyomavirus 7, Human polyomavirus 10, Human polyomavirus 11, and Human polyomavirus 14.*
12. The method according to any one of the previous claims, wherein the virus is *Human Herpesvirus 5* (also known as human cytomegalovirus), *Murid Herpesvirus 1* (also known as murine cytomegalovirus), or *Simian virus 40*.
13. Use of DNA- bead based array in any one of the claims 1 to 12 to determine:
- (a) the presence of integrated viral DNA in a mammalian cell; and
- (b) methylation status of at least one CpG site of the virus,
- thereby determining the status of the virus.
14. Use of claim 13, wherein the mammalian cell is an immortalised cell line derived from the mammal.
15. DNA bead-based array for use in the method of any one of the claims 1 to 12, for determining
- (a) the presence of integrated viral DNA in a mammalian cell; and
- (b) methylation status of at least one CpG site of the virus,
- thereby determining the status of the virus.