

A METHOD OF CLASSIFYING CHO CELLS

FIELD OF THE INVENTION

5 The invention relates to a method based on epigenetics for quantitatively and qualitatively classifying a CHO cell based on phenotypes of interest for use in bioprocessing. In particular, the method measures the methylation levels of specific CpG sites and using a specific statistical algorithm determines the phenotypes of interest expressed by the CHO cell and based on the phenotypes of interest expressed by the cell, classifies the CHO cell.

10

BACKGROUND OF THE ART

Chinese Hamster Ovary (CHO) cells are known to be the workhorses for the industrial production of recombinant therapeutic proteins since 1987 and are hence widely used for biologics production. About 15 70% of all recombinant biopharmaceutical proteins and all monoclonal antibodies approved since 2016 are being manufactured in CHO cells. Several advantages of utilizing CHO for biologics production include tolerance to genetic manipulations, ease of adaptation to manufacturing process scales, rapid growth rates, and ability to perform human-compatible post-translational modifications. However, the biologics production system in CHO faces a bottleneck due to the loss of protein productivity over time.

20

Initial expression and /or production of proteins by a cell line is often high; however, the production reduces during prolonged culture. This results in decreased process yield, impacts timelines and increases costs. Changes in cell culture environment can result in an alteration of cell behaviour and protein productivity of the producer cell line.

25

The current methods of determining the suitability of a CHO clone for target protein production are not only time-consuming, but also not specific enough as readout for the interplay between the cell and its immediate environment, which is critical for selection of clones or cells for optimal protein production. For example, US2017/0081732 discloses a method for the selection of a long-term polypeptide expressing or 30 secreting cell based on the determination of histone acylation which is measured using PCR. However, the method is complicated and not accurate. Further, PCR being a non-high throughput method does not allow for examination of genome wide CpGs, making the process incomplete.

Further, genetically identical CHO clones can still result in heterogenous phenotypes, creating instability 35 to an established process, inefficiency and financial loss during heterologous protein production at an industrial scale. Methods to compare and select CHO clones that use only phenotypic analyses are not able to guarantee consistency over time. Genotype comparisons of CHO clones cannot define how genes are expressed differentially to adapt to environmental conditions. As shown by Wippermann A, et al., Appl Microbiol Biotechnol. 2014 Jan;98(2):579-89, supplementation of butyrate which is known to enhance cell 40 specific productivities in CHO cells also led to alterations of epigenetic silencing events.

Accordingly, there is a need in the art for a tool that is efficient and affordable to globally evaluate and regulate CHO metabolism and protein production. There is also a need in the art for methods of selection and maintenance of identical CHO populations in order to improve speed, quality, efficiency and consistency of production. In particular, there is a need in the art for new descriptive and predictive markers for positive and valuable phenotypes in CHO cells that can be used to identify and classify CHO cells early that are well suited for genetic modifications and/or heterologous protein production.

DESCRIPTION OF THE INVENTION

10

The present invention solves the problems above by providing an accurate, simple and reproducible means of classifying CHO cells that are fit for growth, and protein production at an early stage of the bioprocessing. In particular, the present invention provides a means of ensuring, stability, efficiency and reduction of financial loss during heterologous protein production particularly at an industrial scale.

15

The inventors have identified a method of classifying CHO cells by using epigenetics, where DNA methylation of the CHO cell's genome is measured and the level of DNA methylation is correlated with phenotypes of interest expressed by the cells using a machine learning based model. That is, measuring DNA methylation at specific locations (CpG sites) enables making accurate predictions of the expression of phenotypes of interest in any CHO cell tested and this enables the cell to be classified. Since genotype comparisons of CHO clones cannot define how genes are expressed differentially to adapt to environmental conditions, and phenotypic analyses alone are not able to guarantee consistency over time, epigenetic methods, specifically DNA methylation therefore provides a state-of-the-art technology to determine at any stage of bioprocessing of a CHO cell, particularly at an early stage of biologics production in the CHO cell, whether the CHO cell tested shows any, and if so which, phenotypes of interest to determine its suitability for genetic modification and other science research. The method according to any aspect of the present invention allow the use of DNA methylation as a tool to classify CHO cells based on protein production quantitatively and qualitatively.

30 According to one aspect of the present invention, there is provided a method of establishing a biomarker for a first phenotype of interest of a CHO cell, the method comprising the steps of:

- (a) measuring methylation values of all CpG sites within genomic DNA obtained from a population of CHO cells that are a representation of the first phenotype of interest and are part of the training samples,
- 35 (b) defining a set of specific CpG sites having consistent and reproducible methylation values in the training samples of step (a); and
- (c) processing the methylation values of step (b) and phenotypes of interest of the training samples using a machine learning based model;

thereby obtaining the specific CpG sites with corresponding weighting factors defining the biomarker for
40 the first phenotype of interest.

In particular, the machine learning based model is a classifier algorithm (Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Stanford, CA: Stanford University.) and may be selected from the group consisting of Random
5 forest, decision trees (A. Floares and A. Birlutiu, The 2012 International Joint Conference on Neural
Networks (IJCNN), Brisbane, QLD, Australia, 2012, pp. 1-7 , Support Vector Machines(SVM) (Wang, C.,
et al. Scientific Reports, (2020) 10:5880 and Yousef, M. et al, BMC Bioinformatics 2009, 10:337), K
nearest neighbours (KNN) (Wang, C.,2020), neural networks (Daoud, M. et al., 2019, 97: 204-214), multi-
layer perceptron (Wang, C.,2020 and Daoud, M. et al., 2019), and Gaussian mixture models (Prabakaran,
10 I. et al.Cancer Res; 79(13), 2019: 3492-3502).

In one example, the set of specific CpG sites in step (b) may be finally selected using machine learning
techniques, such as random forests (Breiman L (2001). "Random Forests". Machine Learning. 45 (1): 5–
32. doi:10.1023/A:1010933404324). The random forest predictor is a powerful machine learning algorithm
15 that can handle large datasets with high-dimensional features, such as DNA methylation data. It works by
constructing multiple decision trees based on random subsets of the data and features, and then
combining their predictions to obtain a more accurate and robust result. In the context of selecting specific
CpG sites for CHO cell line engineering, the random forest predictor can be trained on a set of labeled
samples with known phenotypes, such as high or low productivity, growth rate, or viability. The input
20 features are the methylation levels of CpG sites across the genome, and the output is the predicted
phenotype of each sample. By analyzing the importance scores of individual CpG sites in the random
forest model, one can identify the most informative and discriminative sites that contribute to the
phenotype prediction.

25 The biomarker according to any aspect of the present invention may also be called a 'CHO methylation
clock' which provides a correlation between phenotypes of interest expressed by a CHO cell and DNA
methylation patterns. Using the method according to any aspect of the present invention, a CHO cell may
be classified based on its methylation pattern and thereby the phenotype of interest it expresses or is
expected to express.

30 The term 'biomarker' as used herein refers to a naturally occurring characteristic by which a particular
pathological or physiological process can be identified in a CHO cell. In particular, the biomarker refers to
a feature, particularly a DNA methylation pattern in a CHO cell that may be used to identify a phenotype
of interest in the CHO cell. The biomarker for identifying a phenotype of interest in the CHO cell thus
35 refers to specific CpG sites with corresponding weighting factors as parameters. The biomarker for a first
phenotype of interest refers to a first set of specific CpG sites with corresponding weighting factors. The
biomarker for a second phenotype of interest refers to a second set of specific CpG sites with
corresponding weighting factors. The biomarker for a third, fourth and subsequent phenotypes of interest
refers to a third, fourth and subsequent set of specific CpG sites with corresponding weighting factors.
40 The biomarkers may be compiled to obtain a collection of biomarkers to determine a range of phenotypes

of interest of a CHO test cell. In particular, the set of specific CpG sites are determined from all the CpG sites of the CHO cell based on the methylation values measured in step (a) using any method known in the art for measuring DNA methylation. In particular, the specific CpG sites are selected based on expressing reliable methylation values (i.e. the methylation values of the specific CpG sites are consistent and reproducible using any method known in the art for measurement of methylation). The set of specific CpG sites result in the same or significantly similar methylation values when DNA methylation is measured using the same method over many more than one measurement. Accordingly, the CpG sites result in consisting and reproducible readings of methylation values.

Steps (b) and (c) according to any aspect of the present invention may be carried out using a computer.

The term 'CHO cell genome' herein refers to the genomic DNA of the CHO cell that excludes the DNA of a virus, particularly CMV and SV40, that are used to introduce foreign DNA to the cell. In particular, the CHO cell genome may denote the cell with a genome make-up that is in a form as seen naturally in the wild. The term may also include genes which have been added to the CHO genome by genetic modification (i.e. with regard to improved production of protein etc.) but not necessarily or not genes and promoters of viruses that have been used to introduce the genes into the CHO genome. The term "CHO cell genome" therefore may exclude virus genes and promoters and/or may include endogenous or homologous genes of the CHO cell and/or genetically modified endogenous or homologous genes of the CHO cell and/or intergenic genes, DNA found between the genes of the CHO cell.

The CHO cell line 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) and Agarabi CHO (ATCC).

As used herein, the term 'population of CHO cells' refers to more than one CHO cell used as reference CHO cells that display at least one phenotype of interest. The different CHO cells expressing the phenotype of interest are then used to determine different CpG sites that are commonly methylated to determine a methylation pattern or reference methylation profile for the first phenotype of interest. In one example, the different CpG sites are collected from more than one CHO cell where each cell displays a first phenotype of interest to obtain a reference methylation profile for a first phenotype of interest or a biomarker for the first phenotype of interest. The reference methylation profile according to any aspect of the present invention may thus not be a naturally occurring methylation profile from a single CHO cell but an artificial profile obtained from combining relevant CpG sites from different reference CHO cell lines, each with at least one phenotype of interest. The population of CHO cells are considered a representation of the first phenotype of interest and the methylation patterns from each of the CHO cells that form the population of CHO cells for the first phenotype of interest form part of the training samples for producing the biomarker for the first phenotype of interest. In one example, the different CpG sites are collected

from a single reference CHO cell that displays at least one phenotype of interest. In another example, the different CpG sites are collected from more than one CHO cell where each cell displays at least one phenotype of interest. The reference profile may be defined by multiple training samples through multivariate statistical methods, such as Principal Component analysis or Multi-Dimensional Scaling.

5 As used herein, the term 'phenotype of interest' in connection with a CHO cell refers to the cell displaying at least one the following characteristics selected from the group consisting of optimal heterologous protein production, phenotypic homogeneity, protein quality, optimal carbohydrate metabolism, optimal amino acid metabolism, optimal lipid metabolism, optimal cell survivability and combinations thereof. In particular, the phenotype of interest refers to a characteristic that the CHO cell according to any aspect of
10 the present invention displays that is beneficial to the survival of the cell, suitability of the cell for protein production and the overall protein production of the cell. All these phenotypes together may be used to classify the CHO cell and determine the cells suitability to be used in protein production. The more phenotypes of interest that the CHO cell is predicted to express or is classified in, the more suitable it is for heterogenous protein production and do be used for larger scale production of proteins.

15 The term 'suitability' as used herein, refers to a CHO cell line that is fit for optimal heterologous protein production. In one example, a CHO cell may be considered suitable for optimal heterologous protein production before a transgene is introduced into the cell. In this case, the CHO cell may have at least one phenotype of interest or characteristics that enable the cell to grow well and allow for easy uptake of the
20 transgene of interest and following the uptake of the transgene, allow for optimal heterologous protein production, where the protein is a product of the transgene of interest. These characteristics or phenotype of interest include at least optimal glucose consumption, growth rate, lactic acid production, ammonia accumulation and the like. When a CHO cell is confirmed of displaying at least one of these phenotypes of interest, the CHO cell may be considered suitable for optimal heterologous protein production when the
25 transgene of interest is introduced into the cell.

In another example, a CHO cell may be considered suitable for optimal heterologous protein production after the transgene has been introduced into the cell. In this case, a CHO cell is genetically modified using methods known in the art to introduce a transgene into the cell and the genetically modified cell is
30 capable of optimal heterologous protein production where the protein is a product of translation of the transgene. The CHO cell in this example, may have a least one phenotype of interest that enables the genetically modified cell line to have good viability and optimal target protein production. These phenotypes of interest may include cell viability (survivability), protein productivity (in terms of protein quantity and quality), phenotypic homogeneity, cell exhaustion, and the like. Accordingly, the method
35 according to any aspect of the present invention may be used on a CHO cell that has been genetically modified (i.e. with transgene introduced into the cell) or on a CHO cell that has not yet been genetically modified. In both cases, the CHO cell may be for use in heterologous protein production.

As used herein, the term 'transgene' refers to a gene that is taken from the genome of one organism and inserted into the genome of another organism by artificial techniques used in genetic modification. For example, a human gene is artificially introduced into the genome of CHO cells for the production of at least one protein of interest, particularly therapeutic proteins.

5

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.

10

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.

15

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.

20

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.

25

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.

30

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 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 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

40

'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
5 productivity, and protein quality, or phenotypic homogeneity, protein productivity, and protein quality.

The term 'protein productivity' as used herein refers to a measure of the amount of protein made per viable cell at a single titre point. It is calculated by dividing the titre (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
10 spectrometry (MS) and the like.

The term 'carbohydrate metabolism', as used herein refers to almost all or all of the biochemical processes responsible for the metabolic formation, breakdown, and interconversion of carbohydrates in cells. It involves multiple pathways such as glycolysis, gluconeogenesis, glycogenolysis, and glycogenesis. For example, glycolysis is one of the key metabolic pathways of CHO cells. Through glycolysis, CHO cells consume glucose as the main carbon source for energy production and generate
15 lactate as the most common metabolic by-product. Particularly, the term 'optimal carbohydrate metabolism' refers to the ideal or best carbohydrate metabolism possible by a CHO cell.

Similarly, the term 'amino acid metabolism' as used herein refer to the whole of the biochemical processes responsible for the metabolic formation, breakdown, and interconversion of amino acids in
25 CHO cells. Amino acids are the basic building blocks of proteins and constitute all proteinaceous material of the cell including the cytoskeleton, protein component of enzymes, receptors, and signalling molecules. In addition, amino acids are utilized for the growth and maintenance of cells. For example, glutaminolysis is a key metabolic pathway of CHO cells. Glutaminolysis is the prevalent pathway through which CHO cells assimilate organic nitrogen for biomass synthesis while releasing ammonium as the main by-
30 product. Particularly, the term 'optimal amino acid metabolism' refers to the ideal or best amino acid metabolism possible by a CHO cell.

The term 'lipid metabolism' as used herein refers to the synthesis and degradation of lipids in CHO cells, involving the breakdown or storage of fats for energy and the synthesis of structural and functional lipids.
35 Lipids are the major component of cellular membranes, act as secondary messengers in cell communication, involved in signalling, transport and secretion. Lipids are also an important source of energy through β -oxidation and the tricarboxylic acid (TCA) cycle. Lipid metabolism can have a significant impact on cell growth. For example, the process of triacylglycerol synthesis and degradation in CHO cells

can greatly affect overall cellular metabolism and viability. Particularly, the term 'optimal lipid metabolism' refers to the ideal or best amino acid metabolism possible by a CHO cell.

Carbohydrate, amino acid and lipid metabolism can be determined by Metabolite Detection Assays, HPLC and bioprocess analyser. These methods are further disclosed at least in Coulet, M. et al., Cells (2022), 11, 1929; Fan Y, et al., Biotechnol Bioeng (2015) 112(3):521–535 and Ali AS, et al., Biotechnol J.(2018); 13(10):e1700745.

In particular, the phenotype of interest according to any aspect of the present invention is selected from the group consisting of phenotypic homogeneity, protein quality, optimal carbohydrate metabolism, optimal amino acid metabolism, optimal lipid metabolism, optimal heterologous protein production, optimal cell survivability and combinations thereof.

In context of the present invention, the term "methylation value" refers to the average methylation value of at least one cytosine (C) residue within the genomic DNA sequence of a CHO cell. Both β -value and M-value may be used as metrics to measure (average) methylation levels. The M-value is more statistically valid for the differential analysis of methylation levels. However, the β -value is much more biologically interpretable and needs to be reported when M-value method is used for conducting differential methylation analysis. Any method known in the art may be used to determine the methylation value of a CpG site or a region in the DNA with many CpG sites. In one example, DNA methylation value is determined using a DNA methylation bead based array.

The term "methylation ratio" refers to number of methylated cytosine(s) divided by the total number of cytosine(s) covered at the specific site(s). The methylation ratios for the CpG sites in step (b) may be advantageously determined using bisulfite sequencing.

The term "read coverage of the CpG site" is to be understood as the number of reads that align with the known CpG site in the reference sequence. The methylation ratio and the read coverage of the genomic CpG sites may be determined in step (a) using bisulfite sequencing.

In one example, in step (a) a methylation ratio and read coverage of the CpG sites are determined using bisulfite sequencing; and in step (b), the set of specific CpG sites are defined using a cutoff value using bisulfite sequencing.

Step b) according to the first aspect of the present invention may include a bisulfite conversion process.

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).

Whole genome bisulfite sequencing is a genome-wide analysis of DNA methylation based on the sodium bisulfite conversion of genomic DNA, which is then sequenced on a next-generation sequencing platform. The sequences are then re-aligned to the reference genome to determine methylation states of the CpG dinucleotides based on mismatches resulting from the conversion of unmethylated cytosines into uracil. In particular, in step (a) a methylation ratio and read coverage of the CpG sites are determined using bisulfite sequencing; and in step (b), the set of specific CpG sites are defined using a cutoff value using bisulfite sequencing.

The term "test" used in conjunction with the term cell herein refers to a cell that is subjected to the method according to any aspect of the present invention and is the basis for an analysis application of the present invention. A 'test cell' is therefore a CHO cell or a group of CHO cells being tested according to any aspect of the present invention, or a profile being obtained or generated in this context. Conversely, the term "reference" or 'control' shall denote, mostly predetermined, entities which are used for a comparison with the test entity. In particular, a 'test cell' refers to a cell being tested for suitability of optimal homologous protein production where the methylation status has to be determined and a 'control' or 'reference' refers to a cell which is known to display optimal homologous protein production or a methylation profile thereof.

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. In some cases a CpG site may not be considered fully hypermethylated or hypomethylated but a value may be given that is a measure of methylation of the CpG site. Accordingly, methylation may be quantified and may not always be an absolute case of hypermethylation or hypomethylation.

5

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

10 The phenotype of interest-correlated training sample of a CHO cell in step (a) may, for example, be taken at any stage of growth of the CHO cell. The sample is always the CHO genomic DNA.

Any machine learning based model may be used to select CpGs, differentially methylated regions (DMRs), lowly methylated regions (LMRs) and CpGs of a subset of that are most predictive of each phenotype of interest. These selected specific CpGs according to any aspect of the present invention can
15 then be used to develop a means of classifying CHO cells into different classes based on phenotypes of interest and DNA methylation signatures.

In particular, the specific CpG sites according to any aspect of the present invention are distributed within low methylated regions (LMRs), CpG islands, variably methylated sites and/or differentially methylated
20 regions (DMRs) in the genome of the CHO cell.

Low Methylated Region (LMR) is a region of the genome wherein less than 60% of CpGs in that region are methylated. More in particular, less than 50%, 40%, 30%, 20% or 10% of the CpGs in the LMRs are methylated. Any method known in the art may be used to identify or detect LMRs in the genomic DNA.
25 Well known methods include using programmes such as MethylSeekR. In particular, LMRs in the genomic DNA have at least three consecutive CpGs and have no single nucleotide polymorphisms (SNPs) in any of the CpG positions. Even more in particular, LMRs in the genomic DNA are identified based on the method disclosed at least in Burger, L., (2013) *Nucleic Acids Research*, 41 (16): e155 and/or Stadler, M., (2011) *Nature* 480, 490–495. LMRs are known to have an average methylation ranging from
30 10% to 50%; are regions of low CG density which do not overlap with CpG islands; tend to be enriched for H3K4me1, DHSs, and p300/CBP; and/or are primarily located distal to promoters in intergenic or intronic regions. In particular, LMRs:

- have an average methylation ranging from 10% to 50%,
- are regions of low CG density;
- 35 - are enriched for Histone H3 monomethylated at lysine 4 (H3K4me1), DNase I hypersensitive sites (DHSs) and transcriptional coactivators CREB binding protein (CPB) and p300;
- are primarily located distal to promoters in intergenic or intronic regions; and/or
- have no single nucleotide polymorphisms (SNPs) in any of the CpG positions.

Low-methylated regions (LMRs) represent a key feature of the dynamic methylome. LMRs are local reductions in the DNA methylation landscape and represent CpG-poor distal regulatory regions that often reflect the binding of transcription factors and other DNA-binding proteins. LMRs were originally described in the mouse (Stadler et al. (2011) *Nature*: 480, 490–95). Evolutionary conservation of LMRs beyond mammals has remained unexplored.

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 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 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.

Differentially methylated regions (DMRs) are genomic regions with different methylation statuses among multiple biological samples like tissues, cells, individuals, etc. These are genomic regions that differ between phenotypes. The statistical power is likely to be greater when adjacent differentially methylated points (DMPs) are considered together as a whole [Gu H et al (2010) *Nat Methods* 2010; 7:133–6]. The lengths of the DMRs may range between a few hundred to a few thousand bases [Rakyan et al (2011) *Nat Rev Genet* 12:529–41, 2011, Bock C (2012) *Nat Rev Genet* 2012; 13:705–19].

DMRs may occur throughout the genome but have been identified particularly around the promoter regions of genes, within the body of genes, and at intergenic regulatory regions. There are two types of regions, predefined or user defined. Regions with special biological meaning, such as CpG islands, CpG shores, UTRs and so on, are predefined. Many traditional statistical testings, including t-test and Wilcoxon rank sum test, can be performed at a region level. For user-defined regions, criteria such as a fixed region length, fixed numbers of significant and adjacent CpG sites, significant and smoothed estimated effect sizes, etc.

Partially methylated domains (PMDs) are extended regions in the genome exhibiting a reduced average DNA methylation level. They cover gene-poor and transcriptionally inactive regions and tend to be heterochromatic.

Differentially methylated Positions (DMP) are CpG sites with different DNA methylation status across different biological samples and regarded as possible functional regions involved in gene transcriptional regulation.

- 5 In particular, the steps (a)-(c) according to any aspect of the present invention are repeated for a second phenotype of interest and subsequent phenotypes of interests thereafter such that there is at least one biomarker for each phenotype of interest to produce a compilation of biomarkers for different phenotypes of interest.
- 10 In particular, the biomarker according to any aspect of the present invention is a set of specific CpG sites, corresponding weighting factors and intercept of the linear model equation that are relevant to a first phenotype of interest. A compilation of biomarkers therefore refers to more than one biomarker wherein each biomarker is correlated to at least one phenotype of interest displayed by the CHO cell.
- 15 The coverage cutoff value defined in step (b) may be a minimum of 10, 9, 8, 7, 6, 5, 4 or 3. In particular, the coverage cutoff value defined in step (b) may be a minimum of 3.

According to a further aspect of the present invention, there is provided a method of classifying a CHO test cell, the method comprising the steps of:

- 20 (a) measuring DNA methylation levels of specific CpG sites in extracted genomic DNA from the CHO test cell; and
- (b) comparing the methylation levels of these specific CpG sites from step (a) with the methylation levels of the same specific CpG sites from reference CHO cells known to express at least one phenotype of interest; and
- 25 (c) deducing therefrom the phenotypes of interest(s) expressed by the test CHO cell, and wherein the specific CpG sites, and the respective weighing factors are parameters of a biomarker for a phenotype of interest, and the classification of the CHO test cell is determined by the phenotype of interest that the CHO test cell expresses.

30

In particular, the biomarkers according to any aspect of the present invention are determined using the method of the first aspect of the present invention.

- The DNA methylation level according to any aspect of the present invention may be determined using any method known in the art. In one example, the methylation levels can be measured using the commercial
- 35 Illumina™ platform. In another example, the methylation levels can be determined using a DNA methylation bead based array.

- Arrays allow for a high-throughput and robust method to determine semi-quantitative/quantitative DNA-
- 40 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

5 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.

10 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 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

15 sites.

The term "array" as used herein refers to an intentionally created collection of probe molecules 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.

20 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

25 according to any aspect of the present invention include slide arrays, silicon 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.

30 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 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

35 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 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 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.

- 5 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.

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
10 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 discloses these techniques in more detail.

15 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 present invention can also be in a fluid format such as a fluid stream of a flow cytometer or similar device. Commercially
20 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 refer to a material or group of materials having a rigid or semi-rigid surface or surfaces. In many examples, at
25 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.

The array or microarray according to any aspect of the present invention may be a very high-density
30 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.

35 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 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 or for the methylated allele of a non-CpG site, the non-methylated allele of the non-CpG site or both.

5

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 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-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 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 probe and the target over the length of the probe.

20

25

According to one aspect of the present invention, there is provided an *in vitro* method for predicting the classification of a CHO test cell, the method comprising the steps of:

- a) measuring DNA methylation levels of specific CpG sites in extracted genomic DNA from the CHO test cell, and,
- b) using a machine learning based model to analyse the methylation levels obtained in step (a), to predict amount of expression of a phenotype of interest in the CHO test cell,

30

wherein the specific CpG sites, are parameters of a biomarker for a phenotype of interest and the classification of the CHO test cell is determined by using the method according to any aspect of the present invention, namely the first aspect.

35

According to another aspect of the present invention, there is provided an *in vitro* method for predicting the classification of a CHO test cell, the method comprising the steps of:

- a) measuring DNA methylation levels of specific CpG sites in extracted genomic DNA from the CHO test cell and multiplying same with their respective weighing factors to obtain the weighted methylation ratios of those CpG sites,

40

- b) using a machine learning based model to analyse the weighted methylation ratios obtained in step (a), to predict amount of expression of a phenotype of interest in the CHO test cell, wherein the specific CpG sites, and the respective weighing factors are parameters of a biomarker for a phenotype of interest and
- 5 the classification of the CHO test cell is determined by using the method according to any aspect of the present invention, namely the first aspect.

- According to yet another aspect of the present invention, there is provided a computer-implemented method of establishing a biomarker for a first phenotype of interest of a population of CHO cells, the
- 10 method comprising the steps of:
- (a) inputting methylation values of all CpG sites within genomic DNA obtained from a population of CHO cells that express a specific phenotype of interest associated to cell fitness of the cell and are part of the training samples,
 - (b) identifying and determining specific CpG sites from all the CpG sites in step (a) showing
 - 15 consistent and reproducible methylation values, and
 - (c) correlating the CpG methylation levels of the CpG sites obtained in step (a) with the phenotype of interest using machine learning based model,
- thereby obtaining the specific CpG sites with corresponding weighting factors as parameters defining the biomarker for the first phenotype of interest.
- 20 In particular, the machine learning based model is a classifier algorithm and may be selected from the group consisting of Random forest, decision trees, Support Vector Machines(SVM), K nearest neighbours (KNN), neural networks, multi-layer perceptron, and Gaussian mixture models.

- According to another aspect of the present invention, there is provided a computer program loaded into a
- 25 memory of a computer, implementing the method according to any aspect of the present invention.

- According to yet another aspect of the present invention, there is provided a tangible computer-readable medium comprising computer-readable code that, when executed by a computer, causes the computer to perform the method according to any aspect of the present invention.

- 30 According to a further aspect of the present invention, there is provided a use of the medium according to any aspect of the present invention to predict the classification of a CHO test cell, wherein the classification of the cell is based on the expression of at least one phenotype of interest by the cell.

- 35 Unless stated otherwise, all percentages (%) given are percentages by mass.

- The examples adduced hereinafter describe the present invention by way of example, without any intention that the invention, the scope of application of which is apparent from the entirety of the description and the claims, be restricted to the embodiments specified in the examples.

- 40

BRIEF DESCRIPTION OF FIGURES:

Figure 1 is a graph showing IgG productivity being categorized into 3 groups based on the productivity level with an equal number of samples: low, medium and high productivity groups.

- 5 Figure 2 is a graph of Receiver Operating Characteristic (ROC) curves using Support Vector Machine (SVM) for the 3 productivity groups.

EXAMPLES:**EXAMPLE 1:**

- 10 *Predicting the specific productivity (pg/cell/day) from CHO cells*

Wet-Lab methodology

- For this experiment, sixty transgenic CHO clones are grown in basal at 37°C, 8% CO₂, at a shaking speed of 225 RPM. The sixty transgenic CHO cell lines include low producers (specific productivity <10pg/cell/day), intermediate producer (specific productivity 10-20pg/cell/day) and high producers (specific productivity >20pg/cell/day). The flasks are seeded with 3E5 viable cells/mL on day 0 and the culture is fed with appropriate feed on Day 3, 5, 7, 9, 11 and glucose is topped up to 6g/l using 45% glucose when it drops below 2g/l. The fed-Batch culture of 60 clones are maintained for 14 days. Cell count, cell viability, and specific productivity are measured for day 14 and cell pellets are collected on day 14.

20

DNA Extraction

- DNA is extracted using the PureLink Genomic DNA Isolation Minikit kit (Invitrogen), including RNAase treatment following the manufacturer's instructions. DNA quantity is measured by PicoGreen assay and DNA quality is assessed via NanoDrop (Thermo Scientific) to ensure the A₂₆₀/A₂₈₀ ratio is ≤ 1.8. A small amount of sample is then also analysed using automated electrophoresis on TapeStation (Agilent) to ensure each sample contains high molecular weight.

25

Sequencing Analysis

- The genomic DNA (500ng) from the samples are used to prepare libraries for Whole Genome Bisulfite Sequencing (WGBS). The sequencing of the libraries are performed by a third party on a NovaSeq platform which generated 125GB data per sample with 20X coverage.

30

Bioinformatics methodology

- 35 Preprocessing of Whole-genome bisulfite sequencing (WGBS) data

Raw WGBS data undergoes quality control followed by adaptor trimming with trim galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The paired reads are then mapped to the reference genome of [CriGri-PICRH-1.0] (https://www.ncbi.nlm.nih.gov/assembly/GCF_003668045.3/) by Bismark (<https://www.bioinformatics.babraham.ac.uk/projects/bismark/>). Bismark is also utilized to

perform deduplication and extract the number of reads having methylated CpGs and the number of total reads at this position. Methylation ratios are determined by dividing the number of methylated reads by total reads.

5 Preprocessing of methylation array data

The customized chip array data processing is 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 are defined as methylated signal/(methylated signal + unmethylated signal). It 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.

Model establishment

15 A classifier is built to predict phenotype based on the methylation ratio of each CpG. The classifiers are based on random forest, . Hyperparameters for the, random forest model, number of trees, maximum depth, are tuned through cross-validation. Prediction scoring metrics, for example, accuracy or area under the receiver operating characteristic curve (ROC AUC) are used to evaluate the model performance for determining optimal hyperparameters. With fine-tuned hyperparameters, the classifier
20 was fitted to the full dataset. Coefficients or feature importance of the features are analysed to understand the impact on phenotype.

EXAMPLE 2:

Predicting the growth status in the form viable cell density (cells/ml) from CHO cells

25

Wet-Lab methodology

For this experiment, sixty transgenic CHO clones are grown in basal medium supplemented at 37°C, 8% CO₂, at a shaking speed of 225 RPM. The sixty transgenic CHO cell lines include slow growers (viable cell density <1E7cells/ml), intermediate growers (viable cell density 1E7-3E7 cells/ml) and fast growers
30 (viable cell density >3E7 cells/ml). The flasks are seeded with 3E5 viable cells/mL on day 0 and the culture is fed with appropriate feed on Day 3, 5, 7, 9, 11 and glucose is topped up to 6g/l using 45% glucose when dropped below 2g/l. The fed-Batch culture of 60 clones is maintained for 14 days. Cell count and cell viability are measured for day 7 and cell pellets are collected on day 7.

The same steps of DNA Extraction, Sequencing Analysis and bioinformatics methodology in Example 1
35 was used to predict the growth status in the form viable cell density (cells/ml) from CHO cells tested.

EXAMPLE 3

CHO cell culture

The cells used in this study belong to the Humira431 CHO cell line (A*Star BTI), which has been derived from CHO DG44 cells modified to express a therapeutic antibody (Adalimumab biosimilar). Cells were seeded at 3×10^5 cells/ml in 30 ml media in 125 ml shake flasks and culture conditions maintained at 37°C, 8% CO₂, with shaking at 150 rpm. For fed-batch cultures, 10% v/v of EX-CELL Advanced CHO Feed 1 without glucose (24368C, Merck) was added every alternate day from days 3 to 14, and D-(+)-Glucose solution (G8769, Merck) added when glucose concentration fell to 2 g/L. Samples were taken from each flask every day from day 3 until the end of the culture for cell count and metabolite analysis. Viable cell density (VCD) was determined using the Vi-CELL BLU analyser (Beckman Coulter) and metabolite concentrations were measured using the Cedex Bio Analyser (Roche).

10 *Whole genome bisulfite sequencing (WGBS):*

Humira CHO samples were cultured either by fed-batch culture or by batch culture, the IgG titre were measured, followed by whole genome bisulfite sequencing (WGBS).

The EZ DNA Methylation-Gold™ Kit was used for bisulfate conversion, while the VAHTS Universal Pro DNA Library Prep Kit was used for library preparation. The Illumina NovaSeq 6000 was used for whole genome bisulfate sequencing (WGBS).

Data processing:

The raw WGBS data underwent quality control with FastQC and adaptor trimming using TrimGalore. The aligned reads were mapped to the CriGri-PICRH-1.0 reference genome using Bismark, which was also used for deduplication and extraction of beta values for each CpG site. Only samples with above 10 million CpG sites with coverage above 10 were retained for the following analysis.

To ensure the methylation data's reliability and suitability for model building, all samples needed to share the same CpG sites with sufficient coverage. Various coverage thresholds were hence explored to determine the optimal level. The threshold of 10 was then selected based on its balance of stringency and the retention of a substantial number of CpG sites. CpGs that associated with culture types were removed to avoid confounding effects, refining the dataset to 543,613 CpG sites in 114 samples suitable for modeling.

IgG productivity was categorized into 3 groups based on the productivity level with an equal number of samples: low, medium and high productivity groups (as shown in Figure 1).

Machine learning model development:

30 Three machine learning classifiers—Random Forest (RF), Logistic Regression, and Support Vector Machine (SVM)—were evaluated to predict outcomes based on the dataset. The implementation and hyperparameter tuning of these models were conducted using the SciKit-Learn library, with GridSearchCV utilized for systematic hyperparameter optimization.

Results:

The performance of three models was evaluated using 5-fold cross-validation, with key metrics such as mean accuracy and average error rate calculated. The results are summarized in Table 1.

5 The Support Vector Machine (SVM) demonstrated the highest mean accuracy (0.792) and the lowest average error rate (0.208) among the models tested. To further evaluate the discriminatory power of the SVM model, the Receiver Operating Characteristic (ROC) curves for each class were plotted, and the corresponding AUC values were computed (Figure 2). Figure 2 indicates that the SVM model provided robust classification performance, with AUC values of 0.98 for the low class, 0.86 for the medium class, and 0.94 for the high class. These results underscore the model's strong capability in distinguishing between different classes.

10 Table 1: model evaluation

	Random forest	Logistic regression	Support vector machine
Mean accuracy	0.562	0.771	0.792
Average error rate	0.438	0.229	0.208
Best error rate	0.304	0.087	0

CLAIMS

1. A method of establishing a biomarker for a first phenotype of interest of a CHO cell, the method comprising the steps of:
 - 5 (a) measuring methylation values of all CpG sites within genomic DNA obtained from a population of CHO cells that are a representation of the first phenotype of interest and are part of the training samples,
 - (b) defining a set of specific CpG sites having consistent and reproducible methylation values in the training samples of step (a); and
 - 10 (c) processing the methylation values of step (b) and phenotypes of interest of the training samples using a machine learning based model;
thereby obtaining the specific CpG sites with corresponding weighting factors defining the biomarker for the first phenotype of interest.
- 15 2. The method according to claim 1, wherein the machine learning based model is a classifier algorithm selected from the group consisting of Random forest, decision trees, Support Vector Machines(SVM), K nearest neighbours (KNN), neuron networks, multi-layer perceptron, and Gaussian mixture models.
- 20 3. The method according to either claim 1 or 2, wherein the steps (b) and (c) are carried out on a computer.
4. The method according to any one of the preceding claims, wherein the steps (a)-(c) are repeated for a second phenotype of interest and subsequent phenotypes of interests thereafter such that
25 there is at least one biomarker for each phenotype of interest to produce a compilation of biomarkers for different phenotypes of interest.
5. The method according to any one of the preceding claims, wherein the DNA methylation value is determined using a DNA methylation bead based array.
- 30 6. The method according to any one of the claims 1 to 4, wherein in step (a) a methylation ratio and read coverage of the CpG sites are determined using bisulfite sequencing; and in step (b), the set of specific CpG sites are defined using a cutoff value using bisulfite sequencing.
- 35 7. The method according to claim 6, wherein the coverage cutoff value defined in step (b) is a minimum of 3.
8. An *in vitro* method for predicting the classification of a CHO test cell, the method comprising the steps of:

- a) measuring DNA methylation levels of specific CpG sites in extracted genomic DNA from the CHO test cell,
- b) using a machine learning based model to analyse the methylation levels obtained in step (a), to predict amount of expression of a phenotype of interest in the CHO test cell,
- 5 wherein the specific CpG sites, are parameters of a biomarker for a phenotype of interest and the classification of the CHO test cell is determined by the phenotype of interests that the CHO test cell expresses, wherein the biomarkers are determined using the method according to any one of the claims 1 to 7.
- 10 9. The method according to any one of the preceding claims, wherein the phenotype of interest is selected from the group consisting of phenotypic homogeneity, protein quality, optimal carbohydrate metabolism, optimal amino acid metabolism, optimal lipid metabolism, optimal heterologous protein production, optimal cell survivability and combinations thereof.
- 15 10. The method according to any one of the preceding claims, wherein the specific CpG sites are distributed within low methylated regions (LMRs), CpG islands, variably methylated sites and/or differentially methylated regions in the genome of the CHO cell.
- 20 11. The method according to any one of the claims 8 to 10, wherein the DNA methylation level in step (a) is determined using a DNA methylation bead based array.
- 25 12. A computer-implemented method of establishing a biomarker for a first phenotype of interest of a population of CHO cells, the method comprising the steps of:
- (a) inputting methylation values of all CpG sites within genomic DNA obtained from a population of CHO cells that express a specific phenotype of interest associated to cell fitness of the cell and are part of the training samples,
- (b) identifying and determining specific CpG sites from all the CpG sites in step (a) showing consistent and reproducible methylation values, and (c) correlating the CpG methylation levels of the CpG sites obtained in step (a) with the phenotype of interest using machine
- 30 learning based model,
- thereby obtaining the specific CpG sites with corresponding weighting factors as parameters defining the biomarker for the first phenotype of interest.
- 35 13. The method according to claim 12, wherein the machine learning based model is a classifier algorithm selected from the group consisting of Random forest, decision trees, Support Vector Machines(SVM), K nearest neighbours (KNN), neural networks, multi-layer perceptron, and Gaussian mixture models.