

PREDICTING THE FITNESS OF A CHO CELL

FIELD OF THE INVENTION

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

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 also 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 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 identifying CHO cells that are fit for growth, and protein production at an early stage of 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 number of CpG (Cytosine-phosphate Guanine) sites in CHO cells' genome for which the level of DNA methylation is correlated with phenotypes of interest expressed by the cells. That is, measuring DNA methylation at these locations (CpG sites) enables making accurate predictions of the expression of phenotypes of interest in any CHO cell tested. 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 the 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 predict protein production quantitatively and qualitatively from CHO cells.

20

25

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:

30

- (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
- (c) performing a penalized regression using the methylation values of step (a) as input and phenotype of interest correlated to the training samples as dependent variable, by applying a penalized regression model;

35

thereby obtaining the specific CpG sites with corresponding weighting factors and intercept of the linear model equation as parameters defining the biomarker for the first phenotype of interest.

40

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 fitness score for an unknown sample may be quantitatively analysed using the set of specific CpG sites related to each phenotype of interest.

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 refers to specific CpG sites with corresponding weighting factors and intercept of the linear model equation as parameters. The biomarker for a first phenotype of interest refers to a first set of specific CpG sites with corresponding weighting factors and intercept of the linear model equation. The biomarker for a second phenotype of interest refers to a second set of specific CpG sites with corresponding weighting factors and intercept of the linear model equation. 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 and intercept of the linear model equation. The biomarkers may be compiled to obtain a collection of biomarkers to determine a range of phenotypes of interest of a CHO 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 by a skilled person 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 different 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 different 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.

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 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 determine a CHO fitness score or the CHO cell fitness level. The more phenotypes of interest or optimal phenotypes of interest that the CHO cell expresses the higher the fitness level and vice versa.

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

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 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 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 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.

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.

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.

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.

5 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
10 various therapeutic proteins.

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
15 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
20 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
25 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).
30

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
35 can be determined using Immunoprecipitation based techniques, Biochemical Assays, Mass 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
40 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 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.

5

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 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.

10 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-product. Particularly, the term 'optimal amino acid metabolism' refers to the ideal or best amino acid metabolism possible by a CHO cell.

15

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. 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.

25 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.

30 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.

35 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 array may be used.

5 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.

10 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.

Step b) according to the first aspect of the present invention may include a bisulfite conversion process. In step c) a regression analysis may be used determine the phenotypes of interest displayed by the CHO cell tested.

15 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
20 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
25 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
30 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.

35 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)).

40 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
5 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
10 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
15 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
20 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.

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.

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.
30

In particular, the penalized regression model is elastic net linear regression model. Elastic net is 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 then be used to develop a model for predicting cell fitness
35 based on DNA methylation signatures. More in particular, the Elastic net algorithm is a machine learning based method that combines traditional Lasso and ridge regression techniques, placing emphasis on model sparsity while appropriately balancing the contributions of correlated variables. It is particularly useful in the method according to any aspect of the present invention for constructing linear models where the number of variables is substantially greater than the number of samples.

40

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 regions (DMRs) in the genome of the CHO cell.

- 5 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. 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 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;
 - 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.

- 30 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.

5 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].

10

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
15 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.

20

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.

25

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.

30

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.

35

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.

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.

The method according to any aspect of the present invention may further comprise a step of (d) optimizing the fit of the regression model such that the number of CpG sites is reduced to a minimum. For optimizing the fit of the regression model in step (d) or in general for optimizing the fit of the regression model, the algorithm preferably applies ridge regression in combination with lasso regression.

- 5 Advantageously, the methods of ridge regression and lasso regression are balanced using an alpha value of between 0 and 1.

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

- 10 (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 methylation levels of the same specific CpG sites from reference CHO cells known to express at least one phenotype of interest; and
- 15 (c) deducing therefrom the phenotypes of interest(s) expressed by the test CHO cell, and wherein the cell fitness of the CHO test cell is determined by the number of phenotypes of interests that the CHO test cell expresses, and wherein the specific CpG sites in steps (a) and (b) are determined using the method according to any aspect of the present invention.

20

The term 'cell fitness' of a CHO cell as used herein refers to the ability of the CHO cell to grow, produce protein, survive and display other positive characteristics that make the cell suitable for use in cell culture. In particular, cell fitness of a CHO cell may be determined by the phenotypes of interest that the cell displays.

25

The phenotype of interest-correlated reference sample in step (b) serves as a control and represents an average methylation level for a pre-determined and specific phenotype of interest in a CHO cell. The method for predicting cell fitness of a CHO cell may be used for testing individual CHO cell and for testing a complete population of CHO cells.

30

The biological sample material deriving from the CHO cell refers to genomic DNA. As used herein, the term "genomic material" refers to nucleic acid molecules or fragments of the genome of the CHO cells. 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 CHO cells.

35

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

- 40 According to a further aspect of the present invention, there is provided an in vitro method for predicting cell fitness 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,
- b) computing the sum over the weighted methylation ratios obtained in step (a) and adding the respective intercept of linear model equation, thus predicting amount of expression of a phenotype of interest in the CHO test cell,

wherein the specific CpG sites, the respective weighing factors and the respective intercept of linear model equation are the parameters of a biomarker for a phenotype of interest and the cell fitness of the CHO test cell is determined by the number of phenotypes of interests that the CHO test cell expresses.

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 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-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 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 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.

5 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 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.

15 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 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).

20 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. 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.

30 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 discloses these techniques in more detail.

40 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 available fluid formats for distinguishing beads include, for example, those used in XMAP(TM) technologies from Luminex or MPSS(TM) methods from Lynx Therapeutics.

- 5 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 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.

10

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

15 on the array.

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

20 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

25 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 in their unaltered state or as aggregates. Targets may be attached, covalently or noncovalently, to a binding

30 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.

35 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

40 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.

According to 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 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 elastic Net linear regression model
- thereby obtaining the specific CpG sites with corresponding weighting factors and intercept of the linear model equation as parameters defining the biomarker for the first phenotype of interest.

According to another aspect of the present invention, there is provided a computer program loaded into a 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.

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 cell fitness of a CHO test cell, wherein the cell fitness of the cell is based on the expression of at least one phenotype of interest by the cell.

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.

BRIEF DESCRIPTION OF FIGURES

Figure 1 is a scatter plot of predicted IgG vs actual IgG on the test datasets from Example 3.

EXAMPLES

EXAMPLE 1:

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

5 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
10 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 is maintained for 14 days. Cell count, cell viability, and specific productivity are measured for day 14 and cell pellets are collected on day 14.

15 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
20 ensure each sample contains high molecular weight.

Whole-genome bisulfite sequencing (WGBS)

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

Bioinformatics methodology

Preprocessing of Whole-genome bisulfite sequencing (WGBS) data

Raw WGBS data undergoes quality control followed by adaptor trimming with trim galore
30 (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
35 total reads.

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
40 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.

5

Model establishment

A penalized regression model (<https://cran.r-project.org/web/packages/glmnet/index.html> in R and <https://scikit-learn.org/> in Python) are used to regress the specific productivity to the methylation ratio of each CpG. Briefly, Hyperparameters, including the alpha values, which control the balance between L1 and L2 regularization, were tuned through cross-validation. Mean Squared Error (MSE), or Root Mean Squared Error (RMSE), or R-squared (R^2) are used to evaluate the model performance for determining optimal hyperparameters. With fine-tuned hyperparameters, the elastic net model is fitted to the full dataset. Coefficients of the features are analyzed to understand the impact on productivity.

15 **EXAMPLE 2:**

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

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 (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 it drops 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 were collected on day 7.

25

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 A260/280 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.

30

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.

35

Bioinformatics methodology

Preprocessing of Whole-genome bisulfite sequencing (WGBS) data

40

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.

Preprocessing of methylation array data

- 10 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 is 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) is 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.
- 15 Background subtraction based on normal-exponential deconvolution using out-of-band probes noob (Triche et al. 2013) and optionally with extra bleed-through subtraction are also implemented.

Model establishment

- A penalized regression model (<https://cran.r-project.org/web/packages/glmnet/index.html> in R and <https://scikit-learn.org/> in Python) is used to regress the cell counts to the methylation ratio of each CpG. Briefly, Hyperparameters, including the alpha values, which control the balance between L1 and L2 regularization, were tuned through cross-validation. Mean Squared Error (MSE), or Root Mean Squared Error (RMSE), or R-squared (R^2) were used to evaluate the model performance for determining optimal hyperparameters. With fine-tuned hyperparameters, the elastic net model was fitted to the full dataset.
- 25 Coefficients of the features were analyzed to understand the impact on cell viability.

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).

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
5 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
10 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
15 to avoid confounding effects, refining the dataset to 543,613 CpG sites in 114 samples suitable for modeling.

Machine learning model development:

Elastic Net penalized regression model was used to predict IgG values based on methylation data from 543,613 CpG sites in 114 CHO Humira samples. The performance of the predictive models was assessed
20 using Root Mean Squared Error (RMSE), a metric that measures the average magnitude of error between the predicted and actual values, with lower RMSE indicating more accurate predictions. Model implementation and hyperparameter optimization were performed using the SciKit-Learn library and GridSearchCV.

Results:

25 The Elastic Net model identified 850 CpG sites (Table 1) that potentially influence IgG productivity through their methylation patterns. The optimal value for the alpha parameter was determined to be 0.1, resulting in a best cross-validation RMSE score of 12.89. The model was then applied to the test datasets to predict IgG values, and the correlation between the predicted and actual values is depicted in Figure 1.

Table 1 850 CpG sites identified using Elastic Net model that potentially influence IgG productivity
30 through their methylation patterns

SR.No.	chromosome	Position	Coefficient	SR.No.	chromosome	Position	Coefficient
0	NC_048595.1	35044432	-0.097858798	51	NC_048595.1	5190577	-0.023656433
1	NC_048595.1	5703465	-0.043595648	52	NC_048595.1	1187698	-0.0295011267
2	NC_048595.1	6080266	-0.07250116	53	NC_048595.1	31647739	-0.0293011847
3	NC_048595.1	64539635	-0.070382019	54	NC_048595.1	17887455	-0.029137448
4	NC_048595.1	5228272	-0.069532281	55	NC_048595.1	12251611	-0.0289461446
5	NC_048595.1	61839731	-0.068262933	56	NC_048595.1	1049051	-0.0285829316
6	NC_048595.1	5414879	-0.057344666	57	NC_048595.1	445118649	-0.0284307112
7	NC_048595.1	5375586	-0.057302397	58	NC_048595.1	2618956	-0.028412788
8	NC_048595.1	452707269	-0.0534345778	59	NC_048595.1	5528620	-0.028148691
9	NC_048595.1	5451140	-0.053300171	60	NC_048595.1	38756530	-0.02783915
10	NC_048595.1	5685178	-0.053061379	61	NC_048595.1	69520536	-0.027721097
11	NC_048595.1	5235557	-0.051249694	62	NW_0232768	55969398	-0.027618766
12	NC_048595.1	2286542	-0.050862701	63	NC_048595.1	441775804	-0.027375474
13	NC_048595.1	12200971	-0.050010768	64	NC_048595.1	245704844	-0.0272796839
14	NC_048595.1	5550915	-0.0489317955	65	NC_048595.1	8278610	-0.027214333
15	NC_048595.1	8305211	-0.048595951	66	NC_048595.1	76203614	-0.026883478
16	NC_048595.1	7608664	-0.04792742	67	NC_048595.1	85987357	-0.0265646974
17	NC_048595.1	17902617	-0.046702984	68	NC_048595.1	15522189	-0.026579397
18	NC_048595.1	6333213	-0.046265656	69	NC_048595.1	10304831	-0.026334468
19	NC_048595.1	5980400	-0.045616168	70	NC_048597.1	75603965	-0.025724626
20	NC_048595.1	64375176	-0.045441989	71	NC_048595.1	53104612	-0.02570295
21	NC_048595.1	5406091	-0.044719566	72	NC_048595.1	15864459	-0.025544119
22	NC_048595.1	8428957	-0.044095779	73	NC_048596.1	665728007	-0.025311558
23	NC_048595.1	5590433	-0.041939443	74	NC_048595.1	456984165	-0.025078795
24	NC_048595.1	5434622	-0.041610263	75	NC_048595.1	50335551	-0.024928687
25	NC_048595.1	5106871	-0.040300416	76	NC_048595.1	454256293	-0.024605528
26	NC_048595.1	7085452	-0.039180147	77	NC_048595.1	17833912	-0.023807399
27	NC_048595.1	4816439	-0.038744105	78	NC_048595.1	16626391	-0.023448972
28	NC_048595.1	23515532	-0.038665448	79	NC_048595.1	164184497	-0.023221207
29	NC_048595.1	24600395	-0.038525767	80	NC_048595.1	5177897	-0.02318703
30	NC_048595.1	5526925	-0.0380565983	81	NW_0232768	244545618	-0.02288578
31	NC_048595.1	15379114	-0.035711604	82	NC_048595.1	452073787	-0.022737369
32	NC_048595.1	5132282	-0.035363903	83	NC_048595.1	53100884	-0.022631325
33	NC_048595.1	5327841	-0.034893223	84	NC_048597.1	15118749	-0.022497229
34	NC_048595.1	9670654	-0.034174149	85	NC_048595.1	16924535	-0.022335466
35	NC_048595.1	6198085	-0.0340338779	86	NC_048595.1	17164296	-0.0222302553
36	NC_048595.1	15108645	-0.0337872934	87	NC_048595.1	259581186	-0.0222287413
37	NC_048595.1	20915840	-0.033672934	88	NC_048595.1	57998625	-0.021909841
38	NC_048595.1	45316840	-0.033280353	89	NC_048595.1	57998625	-0.021909841
39	NC_048595.1	453168405	-0.032803533	90	NC_048595.1	57998625	-0.021909841
40	NC_048595.1	21916648	-0.032425293	91	NC_048595.1	66267117	-0.021713314
41	NC_048595.1	50183753	-0.03214296	92	NC_048595.1	71486863	-0.021173314
42	NC_048595.1	10696453	-0.03124294	93	NC_048595.1	41277466	-0.021113037
43	NC_048595.1	15130684	-0.031140351	94	NC_048595.1	51377693	-0.021056693
44	NC_048595.1	25800957	-0.031339011	95	NC_048595.1	25353129	-0.020902861
45	NC_048595.1	95472077	-0.031324647	96	NC_048595.1	10640931	-0.020825483
46	NC_048595.1	90370259	-0.031155073	97	NC_048595.1	15885230	-0.020252057
47	NC_048595.1	9458726	-0.031112477	98	NC_048595.1	15885230	-0.020252057
48	NC_048595.1	16615548	-0.030759863	99	NC_048595.1	32242731	-0.020166272
49	NC_048595.1	15840795	-0.030547768	100	NC_048595.1	11508432	-0.0198034029

202300097 Foreign Countries

Sr No.	chromosome	Position	Coefficient	Sr No.	chromosome	Position	Coefficient	Sr No.	chromosome	Position	Coefficient
201	NC_048595.1	12801338	-0.009266232	251	NC_048603.1	21796510	-0.006373	301	NW_0232774	112245	-0.004142075
202	NW_0232768	377961	-0.009170903	252	NC_048602.1	11742015	-0.006365886	302	NC_048595.1	5571729	-0.004101628
203	NC_048595.1	19588041	-0.009141753	253	NC_048595.1	81741815	-0.006206387	303	NC_048595.1	10817286	-0.004045639
204	NC_048595.1	83036144	-0.009139013	254	NC_048595.1	5530340	-0.006143813	304	NC_048595.1	55698297	-0.004022062
205	NC_048595.1	5436023	-0.009063372	255	NC_048595.1	439682118	-0.006147567	305	NC_048595.1	13556194	-0.004004168
206	NC_048595.1	16897606	-0.00898321	256	NC_048595.1	4070152	-0.006106585	306	NC_048595.1	48953927	-0.003987622
207	NW_0232774	102908	-0.008883981	257	NC_048595.1	5446347	-0.006089702	307	NC_048595.1	183973595	-0.003925686
208	NC_048595.1	1818610	-0.008880969	258	NC_048595.1	34180236	-0.006080458	308	NC_048595.1	33058528	-0.00382701
209	NC_048595.1	445408666	-0.008828218	259	NC_048595.1	4927306	-0.006018096	309	NC_048595.1	5135726	-0.00379832
210	NC_048595.1	36924382	-0.008768832	260	NC_048595.1	26095803	-0.006014967	310	NC_048595.1	53426046	-0.003725519
211	NC_048595.1	18401357	-0.008695722	261	NC_048595.1	67765997	-0.006007622	311	NC_048595.1	5316942	-0.003628718
212	NC_048595.1	440828532	-0.008628225	262	NC_048595.1	17109302	-0.005970306	312	NC_048595.1	63945879	-0.003552735
213	NC_048595.1	6690380	-0.008623028	263	NC_048595.1	34871282	-0.005913329	313	NC_048603.1	23517421	-0.00346377
214	NC_048595.1	4879578	-0.008559992	264	NC_048595.1	45792460	-0.005873795	314	NC_048595.1	24799775	-0.003413949
215	NC_048595.1	11148853	-0.008496716	265	NC_048595.1	8563282	-0.005815282	315	NC_048595.1	41279135	-0.003408255
216	NC_048595.1	36110758	-0.008464031	266	NC_048603.1	22414892	-0.005726363	316	NC_048595.1	44733627	-0.003406584
217	NC_048595.1	32081326	-0.008435541	267	NC_048595.1	58249849	-0.00568503	317	NC_048595.1	34496833	-0.003353461
218	NC_048597.1	109133854	-0.008313108	268	NC_048595.1	87268012	-0.005641864	318	NC_048595.1	63358075	-0.003322211
219	NC_048595.1	88872121	-0.008276556	269	NC_048595.1	28465190	-0.005508715	319	NC_048595.1	34623574	-0.003283166
220	NC_048595.1	13400330	-0.008262883	270	NC_048604.1	106745250	-0.005396354	320	NC_048595.1	6095704	-0.003278349
221	NC_048595.1	271209517	-0.008222448	271	NC_048596.1	3498060	-0.005394632	321	NC_048595.1	25952129	-0.003261567
222	NC_048600.1	119761307	-0.008196359	272	NC_048595.1	4484258	-0.005390911	322	NC_048595.1	16474726	-0.00320408
223	NC_048595.1	24609882	-0.008027438	273	NC_048595.1	16700383	-0.005350156	323	NC_048600.1	134305788	-0.00319837
224	NC_048595.1	44969902	-0.007923211	274	NC_048595.1	15700383	-0.005350156	324	NC_048597.1	7267982	-0.00307699
225	NC_048595.1	70174389	-0.007919609	275	NC_048596.1	235597212	-0.005339806	325	NC_048595.1	34877153	-0.002999359
226	NC_048595.1	58112174	-0.007889338	276	NC_048597.1	137260254	-0.005330871	326	NC_048595.1	5819354	-0.002962345
227	NC_048595.1	5146916	-0.00773381	277	NC_048595.1	27912226	-0.005326478	327	NC_048600.1	120838705	-0.002928676
228	NC_048595.1	66776708	-0.007703896	278	NC_048595.1	88358412	-0.005289485	328	NC_048598.1	153912011	-0.002901258
229	NC_048595.1	88091551	-0.007662324	279	NC_048595.1	16057576	-0.005272206	329	NC_048602.1	17951240	-0.002867218
230	NC_048595.1	44902327	-0.00755257	280	NC_048595.1	268644294	-0.005229786	330	NC_048595.1	21974222	-0.002677699
231	NC_048595.1	20718840	-0.007471027	281	NC_048595.1	26748945	-0.005215164	331	NC_048600.1	12334674	-0.002605091
232	NC_048595.1	32489385	-0.007461232	282	NC_048595.1	11097973	-0.00519454	332	NC_048595.1	77670377	-0.002529689
233	NC_048595.1	5567625	-0.007430509	283	NC_048595.1	51071374	-0.00511521	333	NC_048595.1	32126850	-0.002517188
234	NC_048595.1	4873803	-0.007364778	284	NC_048595.1	15509304	-0.005061207	334	NC_048595.1	32126850	-0.002511964
235	NC_048595.1	5147739	-0.007332697	285	NC_048595.1	5908809	-0.004989119	335	NC_048595.1	25418765	-0.002482033
236	NC_048596.1	46860180	-0.007323919	286	NC_048595.1	22523370	-0.004960032	336	NC_048595.1	4968973	-0.0024212425
237	NC_048595.1	41221261	-0.006994723	287	NC_048595.1	34240090	-0.004935423	337	NC_048595.1	9701396	-0.002385965
238	NC_048595.1	74413805	-0.006862999	288	NC_048595.1	155470317	-0.004876695	338	NC_048595.1	72806415	-0.002318009
239	NC_048595.1	5337075	-0.006830667	289	NC_048595.1	16100403	-0.004842793	339	NC_048595.1	7095808	-0.002317074
240	NC_048595.1	453782828	-0.006827521	290	NC_048595.1	4359707	-0.004882528	340	NC_048597.1	164336236	-0.002286675
241	NC_048596.1	48518284	-0.006870877	291	NC_048595.1	454096410	-0.004880925	341	NC_048595.1	452729724	-0.002266781
242	NW_0232768	7051151	-0.006743962	292	NC_048596.1	50357660	-0.004857806	342	NC_048595.1	27249294	-0.002185652
243	NC_048595.1	9958379	-0.006651058	293	NC_048595.1	53120229	-0.004666983	343	NC_048595.1	16416384	-0.002152285
244	NC_048595.1	5174417	-0.006628131	294	NC_048595.1	73938651	-0.004552556	344	NC_048595.1	68027904	-0.002132091
245	NC_048595.1	287267740	-0.006586409	295	NC_048595.1	92954529	-0.004502614	345	NW_0232768	273468971	-0.002082899
246	NC_048595.1	33334571	-0.006463665	296	NC_048595.1	23881439	-0.00449107	346	NC_048595.1	82274799	-0.002079717
247	NC_048595.1	41739331	-0.006425405	297	NC_048595.1	8140096	-0.004439531	347	NC_048595.1	34633989	-0.002076656
248	NC_048595.1	30504622	-0.006389264	298	NC_048595.1	5336584	-0.004415746	348	NC_048595.1	453792567	-0.001987705
249	NC_048595.1	23665574	-0.006380457	299	NC_048595.1	43141076	-0.004407241	349	NW_0232768	251268827	-0.001913974
250	NC_048595.1	5098421	-0.006360138	300	NC_048595.1	71322759	-0.004404627	350	NW_0232768	251268827	-0.001891475

202300097 Foreign Countries

Sr.No.	chromosome	Position	Coefficient	Sr.No.	chromosome	Position	Coefficient	Sr.No.	chromosome	Position	Coefficient
401	NC_048598.1	15401382	-0.000442472	461	NC_048595.1	32365038	2.48E-05	501	NC_048595.1	29317649	0.000766077
402	NC_048603.1	22534454	-0.000444013	462	NC_048595.1	64380509	2.81E-05	502	NC_048595.1	454665752	0.000846651
403	NC_048595.1	15834050	-0.000433732	463	NC_048595.1	53804473	2.15E-05	503	NC_048595.1	30883943	0.000882375
404	NC_048595.1	15365331	-0.000407433	464	NC_048595.1	75502026	2.25E-05	504	NC_048595.1	36620792	0.000892601
405	NC_048595.1	15304610	-0.000383077	465	NC_048595.1	454029232	2.28E-05	505	NC_048595.1	52945293	0.00089325
406	NC_048595.1	32120167	-0.000388098	466	NC_048595.1	49927501	2.34E-05	506	NC_048595.1	5461663	0.000898
407	NC_048595.1	12005969	-0.000358392	467	NC_048595.1	45158579	2.01E-05	507	NC_048595.1	442488356	0.000899948
408	NC_048595.1	76827660	-0.00031943	468	NC_048595.1	23937479	2.15E-05	508	NC_048595.1	10858521	0.000922245
409	NC_048595.1	461238833	-0.000312408	469	NC_048595.1	26440179	2.18E-05	509	NC_048595.1	14962185	0.000932475
410	NC_048595.1	20394283	-0.000290499	470	NC_048595.1	84646663	2.57E-05	510	NC_048595.1	444720901	0.000939491
411	NC_048595.1	6792543	-0.000289674	471	NC_048595.1	61815636	0.000119631	511	NC_048595.1	47509828	0.001068667
412	NC_048595.1	64013924	-0.000271115	472	NC_048595.1	22501708	0.000119709	512	NC_048595.1	19362495	0.001098934
413	NC_048598.1	184590628	-0.000245639	473	NC_048595.1	9670921	0.000123617	513	NC_048595.1	1868787	0.001123214
414	NC_048595.1	23283769	-0.000243786	474	NC_048595.1	144042767	0.000138501	514	NC_048595.1	53425577	0.001137967
415	NC_048595.1	38872121	-0.000235673	475	NC_048595.1	452942877	0.000152544	515	NC_048595.1	40648888	0.00116652
416	NC_048597.1	162373750	-0.000228622	476	NW_02327686	182893197	0.000154579	516	NC_048595.1	21604044	0.001184638
417	NC_048595.1	87386591	-0.000216381	477	NC_048595.1	39184694	0.000162373	517	NC_048595.1	93472431	0.001232111
418	NC_048595.1	15700645	-0.000203593	478	NC_048595.1	84635943	0.000171817	518	NC_048595.1	33074455	0.001235338
419	NC_048595.1	5309084	-0.000189194	479	NC_048603.1	24542755	0.000174046	519	NC_048597.1	19028231	0.001241633
420	NC_048596.1	50377859	-0.000138468	480	NC_048595.1	18519029	0.000207777	520	NC_048595.1	442921863	0.001260221
421	NC_048603.1	22565117	-0.000138466	481	NC_048595.1	50368691	0.000211288	521	NC_048595.1	7085910	0.001304673
422	NC_048595.1	439652173	-0.000130964	482	NC_048595.1	52641087	0.000221929	522	NC_048595.1	447880483	0.001332655
423	NC_048595.1	51718250	-9.53E-05	483	NC_048595.1	35297415	0.000221993	523	NC_048603.1	220693241	0.00133601
424	NC_048603.1	22176633	-9.47E-05	484	NC_048595.1	96789717	0.000224329	524	NC_048595.1	42071530	0.001380164
425	NC_048595.1	7407769	-8.77E-05	485	NC_048595.1	75829440	0.000231903	525	NC_048595.1	10697337	0.001384188
426	NC_048595.1	34626652	-6.80E-05	486	NC_048595.1	19685936	0.000234046	526	NC_048595.1	51304451	0.001395844
427	NC_048595.1	29028571	-6.80E-05	487	NC_048595.1	40119486	0.000258816	527	NC_048595.1	371994633	0.001424328
428	NC_048595.1	5460832	-6.57E-05	488	NC_048595.1	25369424	0.000286444	528	NC_048595.1	5827866	0.001453272
429	NC_048595.1	24043285	-6.52E-05	489	NC_048595.1	86639151	0.000292695	529	NC_048595.1	8403021	0.001454477
430	NC_048595.1	5443785	-6.52E-05	490	NC_048595.1	26788074	0.000315402	530	NC_048595.1	30552488	0.001517553
431	NC_048600.1	134304441	-5.57E-05	491	NC_048595.1	18837950	0.000333467	531	NC_048595.1	49280801	0.001528862
432	NC_048598.1	61940628	-5.10E-05	492	NC_048595.1	440284791	0.000324285	532	NC_048595.1	30045719	0.001568579
433	NC_048595.1	5667944	-3.45E-05	493	NC_048595.1	61701253	0.000324285	533	NC_048595.1	442152751	0.001590664
434	NC_048595.1	24167544	-3.00E-05	494	NC_048595.1	92285778	0.000333467	534	NC_048595.1	7424244	0.001630732
435	NC_048595.1	61060766	-2.21E-05	495	NC_048595.1	41908043	0.000367965	535	NC_048595.1	33887013	0.00163911
436	NC_048603.1	23507501	-1.75E-05	496	NC_048595.1	34682578	0.000387897	536	NC_048595.1	83485522	0.001699394
437	NC_048595.1	55032911	-1.35E-05	497	NC_048595.1	457174021	0.000398815	537	NC_048595.1	6187792	0.001706709
438	NW_02327696	5545	-1.08E-05	498	NC_048595.1	9600252	0.000437435	538	NC_048595.1	83864942	0.001723983
439	NC_048595.1	31470507	-3.35E-06	499	NC_048595.1	24618090	0.000442023	539	NC_048600.1	124086309	0.001747762
440	NC_048596.1	36779252	-2.11E-06	500	NC_048595.1	92285778	0.000457686	540	NC_048595.1	14152550	0.00181202
441	NC_048596.1	66196290	-1.68E-06	491	NC_048595.1	115151481	0.000463795	541	NC_048595.1	236274753	0.001823965
442	NC_048595.1	43264348	-1.24E-06	492	NC_048595.1	49715127	0.000476276	542	NC_048595.1	14133049	0.001837689
443	NC_048595.1	455436155	2.61E-07	493	NC_048595.1	45042373	0.000477214	543	NW_02327686	128039751	0.001879181
444	NC_048600.1	22687022	2.75E-06	494	NC_048595.1	67002623	0.00048139	544	NC_048595.1	11320746	0.001882538
445	NC_048595.1	456007721	2.75E-06	495	NC_048595.1	26416627	0.000497335	545	NC_048595.1	51884251	0.001880943
446	NC_048595.1	5687527	2.61E-06	496	NC_048595.1	34173940	0.000511387	546	NC_048595.1	15649455	0.001898809
447	NC_048595.1	88281190	2.59E-06	497	NC_048595.1	94788901	0.000619666	547	NC_048595.1	69408158	0.001913417
448	NC_048595.1	28736139	2.11E-05	498	NC_048595.1	16729755	0.000670045	548	NC_048595.1	26699428	0.001969073
449	NC_048595.1	19485094	2.19E-05	499	NC_048595.1	8905038	0.000676129	549	NC_048595.1	16366293	0.001975915
450	NC_048595.1	23800487	2.23E-05	500	NC_048595.1	17684579	0.000700262	550	NC_048595.1	49735986	0.002008871

202300097 Foreign Countries

Sr.No.	chromosome	Position	Coefficient	Sr.No.	chromosome	Position	Coefficient	Sr.No.	chromosome	Position	Coefficient	Sr.No.	chromosome	Position	Coefficient
601	NC_048597.1	166324110	0.004124409	651	NC_048595.1	454401355	0.006701451	701	NC_048595.1	59983701	0.010543709	751	NC_048595.1	37960566	0.016527532
602	NC_048595.1	73039694	0.004151207	652	NC_048598.1	150075555	0.006750183	702	NC_048595.1	27794192	0.010716199	752	NC_048596.1	47057604	0.016550606
603	NC_048595.1	25119357	0.004179554	653	NC_048595.1	5309792	0.006786192	703	NC_048600.1	91844176	0.010760226	753	NC_048595.1	8318001	0.016560289
604	NW_02327773	6917	0.004179808	654	NC_048595.1	45480687	0.006839363	704	NC_048595.1	976331	0.010776331	754	NC_048595.1	16191279	0.016624074
605	NC_048595.1	75115316	0.004200674	655	NC_048595.1	52768136	0.006876144	705	NC_048595.1	9626947	0.011185664	755	NC_048595.1	9701571	0.016695686
606	NC_048595.1	14935938	0.004306518	656	NC_048595.1	195866750	0.006887728	706	NC_048595.1	34087378	0.011185664	756	NC_048602.1	15668883	0.016891939
607	NC_048595.1	31585343	0.004314852	657	NC_048595.1	15287324	0.006902143	707	NC_048595.1	9338656	0.011217273	757	NC_048595.1	17698897	0.016927993
608	NC_048595.1	21452257	0.004355868	658	NC_048595.1	64015880	0.00694152	708	NC_048595.1	23549877	0.011293382	758	NC_048598.1	142523336	0.017175122
609	NC_048595.1	46082801	0.00436182	659	NC_048595.1	116025263	0.007042519	709	NC_048596.1	235229521	0.011352196	759	NC_048596.1	22834028	0.017206701
610	NC_048601.1	24117047	0.004364144	660	NC_048595.1	34051462	0.007113648	710	NC_048595.1	450624286	0.011352196	760	NC_048600.1	89216043	0.017340472
611	NC_048596.1	40048552	0.004413044	661	NC_048595.1	45597046	0.007117347	711	NC_048597.1	85284413	0.011557468	761	NC_048597.1	165543345	0.017627666
612	NC_048595.1	10650436	0.004432964	662	NC_048595.1	10618495	0.007389736	712	NC_048595.1	39037920	0.01163878	762	NC_048595.1	9512041	0.017662974
613	NC_048595.1	27232176	0.004437756	663	NC_048595.1	42413214	0.007432607	713	NC_048595.1	83962022	0.011693374	763	NC_048595.1	63544535	0.017684301
614	NC_048597.1	117387588	0.004455385	664	NC_048595.1	79268819	0.007468517	714	NC_048595.1	57040795	0.01173643	764	NC_048595.1	64207468	0.017850558
615	NC_048595.1	25443013	0.004502866	665	NC_048595.1	89259402	0.007667745	715	NC_048595.1	13568940	0.011743369	765	NC_048595.1	43383650	0.018105255
616	NC_048595.1	10746350	0.004511971	666	NC_048595.1	16559983	0.007720092	716	NC_048595.1	9320817	0.011864909	766	NC_048595.1	26403807	0.018119869
617	NC_048595.1	23520259	0.004535574	667	NC_048595.1	97244520	0.0077720547	717	NC_048595.1	80661527	0.012303137	767	NC_048595.1	445671384	0.018244458
618	NC_048595.1	27798744	0.004589499	668	NC_048595.1	21584643	0.007772761	718	NC_048595.1	23574520	0.012350253	768	NC_048595.1	8236807	0.01826477
619	NC_048595.1	455486471	0.004711672	669	NC_048595.1	19218531	0.007791985	719	NC_048595.1	61798986	0.01236221	769	NC_048595.1	95633173	0.018359597
620	NC_048595.1	457389615	0.004842936	670	NC_048595.1	26544181	0.007932608	720	NC_048595.1	23933540	0.012633745	770	NC_048595.1	9858755	0.018425905
621	NC_048595.1	94878201	0.004907355	671	NC_048595.1	66896470	0.007953054	721	NC_048595.1	440365584	0.012727554	771	NC_048595.1	39621794	0.01857021
622	NC_048595.1	14178938	0.004977343	672	NC_048595.1	37905339	0.008063689	722	NC_048595.1	8832227	0.012949706	772	NC_048595.1	32584601	0.018587711
623	NC_048595.1	390150216	0.005009958	673	NC_048596.1	433977008	0.008271329	723	NC_048595.1	42100554	0.013013029	773	NC_048595.1	22929050	0.018629335
624	NC_048595.1	22891977	0.005017028	674	NC_048595.1	21359869	0.008380369	724	NC_048595.1	66539818	0.013082145	774	NC_048595.1	61059395	0.018696332
625	NC_048595.1	38526047	0.005250325	675	NC_048595.1	28234605	0.008464373	725	NC_048596.1	43644605	0.013161022	775	NC_048595.1	286646345	0.01871746
626	NC_048595.1	38547861	0.005334276	676	NC_048595.1	36357653	0.008606634	726	NC_048595.1	28452673	0.013166732	776	NC_048595.1	89698195	0.018746287
627	NC_048595.1	5697921	0.005386653	677	NC_048595.1	136278561	0.00873152	727	NW_02327681	50264924	0.013294623	777	NC_048595.1	65595659	0.018862657
628	NC_048595.1	442100176	0.005428745	678	NC_048595.1	10435938	0.008773065	728	NC_048601.1	89151850	0.01372091	778	NC_048596.1	37918917	0.019027516
629	NC_048597.1	166323280	0.005480373	679	NC_048599.1	133624211	0.008882515	729	NC_048595.1	37961765	0.013993746	779	NC_048595.1	19263531	0.019112025
630	NC_048595.1	56494803	0.005482062	680	NW_02327681	30201	0.008866593	730	NC_048595.1	20995790	0.014294505	780	NC_048595.1	403814257	0.019471574
631	NC_048600.1	90126710	0.005597667	681	NC_048595.1	4849672	0.009132953	731	NC_048595.1	9671548	0.014370994	781	NC_048595.1	41367791	0.019486955
632	NC_048595.1	346032091	0.005649055	682	NC_048595.1	452998392	0.009314758	732	NC_048595.1	30102866	0.014492354	782	NC_048595.1	9805180	0.019707091
633	NC_048595.1	7987725	0.005779541	683	NC_048595.1	27368376	0.009374632	733	NC_048595.1	9721342	0.014655934	783	NC_048595.1	59612312	0.019815872
634	NC_048597.1	117256261	0.005795385	684	NC_048596.1	38647741	0.00957347	734	NC_048595.1	59318473	0.014657841	784	NC_048595.1	5303560	0.019889215
635	NC_048595.1	9528721	0.005806485	685	NC_048595.1	15427754	0.009647989	735	NC_048595.1	27003638	0.01489589	785	NC_048595.1	9113016	0.020355433
636	NC_048595.1	46970188	0.005916403	686	NC_048595.1	15364964	0.009736055	736	NC_048595.1	88774753	0.01511773	786	NC_048595.1	37813525	0.020514956
637	NC_048595.1	17899616	0.005954728	687	NC_048595.1	151058808	0.009750588	737	NC_048595.1	47854169	0.015209174	787	NC_048595.1	29314661	0.02080633
638	NC_048595.1	5823667	0.006055322	688	NC_048595.1	24976894	0.009816606	738	NC_048595.1	47089063	0.015403096	788	NC_048595.1	9763616	0.021113218
639	NC_048595.1	439667541	0.006088775	689	NC_048595.1	97147979	0.009874988	739	NC_048595.1	23456240	0.01543138	789	NC_048595.1	24150992	0.021249547
640	NC_048595.1	51715903	0.006162918	690	NC_048595.1	60353258	0.009890065	740	NC_048595.1	44752075	0.015484396	790	NC_048595.1	9116297	0.021300559
641	NC_048600.1	21198301	0.006178247	691	NC_048595.1	61669197	0.0098971525	741	NC_048595.1	5022753	0.015813157	791	NC_048595.1	67752965	0.02157108
642	NC_048595.1	39066129	0.006187772	692	NC_048598.1	15264611	0.010002112	742	NC_048595.1	72962989	0.015846841	792	NC_048595.1	42156784	0.021889741
643	NC_048595.1	35857679	0.006286244	693	NC_048596.1	114802668	0.010007291	743	NC_048595.1	30253129	0.015872382	793	NC_048595.1	80729684	0.021986277
644	NC_048595.1	48144675	0.006318766	694	NC_048595.1	15782775	0.010144484	744	NC_048595.1	95694126	0.015976098	794	NC_048595.1	33790539	0.021986693
645	NC_048597.1	15873786	0.0063937	695	NC_048595.1	33501986	0.010187689	745	NC_048600.1	132735763	0.016145049	795	NC_048595.1	82454397	0.022181136
646	NC_048600.1	91269889	0.006415281	696	NC_048595.1	446136067	0.010202686	746	NC_048596.1	48053493	0.01615602	796	NC_048596.1	181761943	0.02221328
647	NC_048595.1	6191965	0.006418455	697	NC_048595.1	62420117	0.010222702	747	NC_048595.1	94497217	0.016227859	797	NC_048595.1	30856615	0.022501682
648	NC_048595.1	6608915	0.006553026	698	NC_048595.1	11529196	0.010491317	748	NC_048595.1	8432243	0.016260136	798	NC_048595.1	456543043	0.022814389
649	NC_048595.1	58837500	0.006574891	699	NC_048595.1	45577715	0.010538865	749	NC_048595.1	14461632	0.016362843	799	NC_048595.1	68351333	0.022881654
650	NC_048598.1	55711164	0.006674095	700	NC_048595.1	68351313	0.01053956	750	NC_048595.1	5141801	0.01643258	800	NC_048595.1	56361873	0.023250275

202300097 Foreign Countries

Sr No.	chromosome	Position	Coefficient
801	NC_048595.1	5166126	0.0233331217
802	NC_048595.1	84578297	0.024090009
803	NC_048595.1	86638928	0.024482464
804	NC_048595.1	55391172	0.024948975
805	NC_048595.1	35584377	0.02541109
806	NC_048595.1	93537485	0.02567699
807	NC_048595.1	23801316	0.025809485
808	NC_048595.1	24037350	0.02600304
809	NC_048595.1	20416537	0.02610551
810	NC_048595.1	352111503	0.026647014
811	NC_048595.1	25188215	0.026811732
812	NC_048595.1	23095146	0.026913662
813	NC_048595.1	13086575	0.027937659
814	NC_048595.1	34856837	0.027947803
815	NC_048595.1	12044036	0.028378518
816	NC_048595.1	7645087	0.029570572
817	NC_048595.1	5300212	0.029637327
818	NC_048595.1	6995306	0.030060736
819	NC_048595.1	27649721	0.030485578
820	NC_048595.1	16701367	0.030825329
821	NC_048595.1	26941100	0.031519144
822	NC_048595.1	28834923	0.032345384
823	NC_048595.1	74137095	0.032699548
824	NC_048595.1	87359456	0.032963779
825	NC_048595.1	25913697	0.033017299
826	NC_048597.1	155032050	0.033205721
827	NC_048595.1	43208067	0.033618036
828	NC_048595.1	53268265	0.033928596
829	NC_048595.1	71379985	0.034308013
830	NC_048595.1	24475811	0.034463702
831	NC_048595.1	40306927	0.03592332
832	NC_048595.1	9716151	0.037713347
833	NC_048595.1	25993256	0.03836288
834	NC_048595.1	459662707	0.039191204
835	NC_048595.1	9747640	0.040089952
836	NC_048595.1	20932565	0.040327315
837	NC_048595.1	9583309	0.04035001
838	NC_048595.1	9476054	0.041469201
839	NC_048595.1	24756892	0.042786009
840	NC_048595.1	37094802	0.043663979
841	NC_048595.1	75294119	0.044262489
842	NC_048595.1	9428083	0.046440998
843	NC_048595.1	10275901	0.0471507
844	NC_048595.1	73001829	0.047745539
845	NC_048595.1	393239755	0.049688028
846	NC_048595.1	5817834	0.051373102
847	NC_048595.1	25942129	0.051942298
848	NC_048595.1	19722723	0.063842599
849	NC_048595.1	41284525	0.070981705

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) performing a penalized regression using the methylation values of step (a) as input and phenotype of interest correlated to the training samples as dependent variable, by applying a penalized regression model;
thereby obtaining the specific CpG sites with corresponding weighting factors and intercept of the linear model equation as parameters defining the biomarker for the first phenotype of interest.
- 15 2. The method according to claim 1, wherein the penalized regression model is elastic net linear regression model.
3. The method according to either claim 1 or 2, wherein steps (b) and (c) are carried out on a
20 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 there is at least one biomarker for each phenotype of interest to produce a compilation of
25 biomarkers for different phenotypes of interest.
5. The method according to any one of the preceding claims, wherein for optimizing the fit of the regression model in step (c), the algorithm applies ridge regression in combination with lasso regression.
- 30 6. The method according to claim 5, wherein the methods of ridge regression and lasso regression are balanced using an alpha value of between 0 and 1.
7. The method according to any one of the preceding claims, wherein the DNA methylation value is
35 determined using a DNA methylation bead based array.
8. The method according to any one of the claims 1 to 6, 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.

9. The method according to claim 8, wherein the read coverage cutoff value defined in step (b) is a minimum of 3.
10. An in vitro method for predicting cell fitness of a CHO test cell, the method comprising the steps of:
- 5 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,
- b) computing the sum over the weighted methylation ratios obtained in step (a) and adding the
- 10 respective intercept of linear model equation, thus predicting amount of expression of a phenotype of interest in the CHO test cell,
- wherein the specific CpG sites, the respective weighing factors and the respective intercept of linear model equation are the parameters of a biomarker for a phenotype of interest and the cell fitness of the CHO test cell is determined by the number of phenotypes of interests that
- 15 the CHO test cell expresses wherein the biomarkers are determined using the method according to any one of the claims 1 to 9.
11. 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
- 20 carbohydrate metabolism, optimal amino acid metabolism, optimal lipid metabolism, optimal heterologous protein production, optimal cell survivability and combinations thereof.
12. 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
- 25 differentially methylated regions in the genome of the CHO cell.
13. The method according to any one of the claims 10 to 12, wherein the DNA methylation level is determined using a DNA methylation bead based array.
14. 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:
- 30 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,
- 35 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 elastic Net linear regression model
- thereby obtaining the specific CpG sites with corresponding weighting factors and intercept of the
- 40 linear model equation as parameters defining the biomarker for the first phenotype of interest.