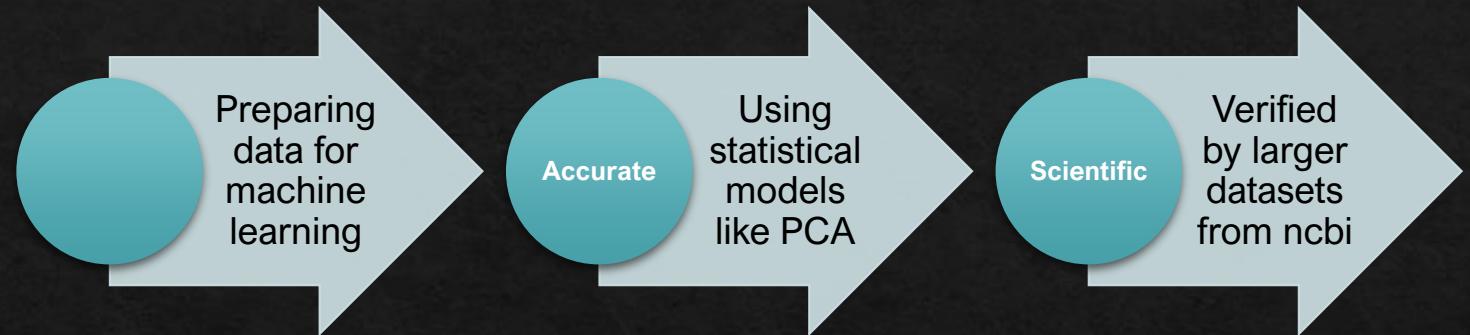


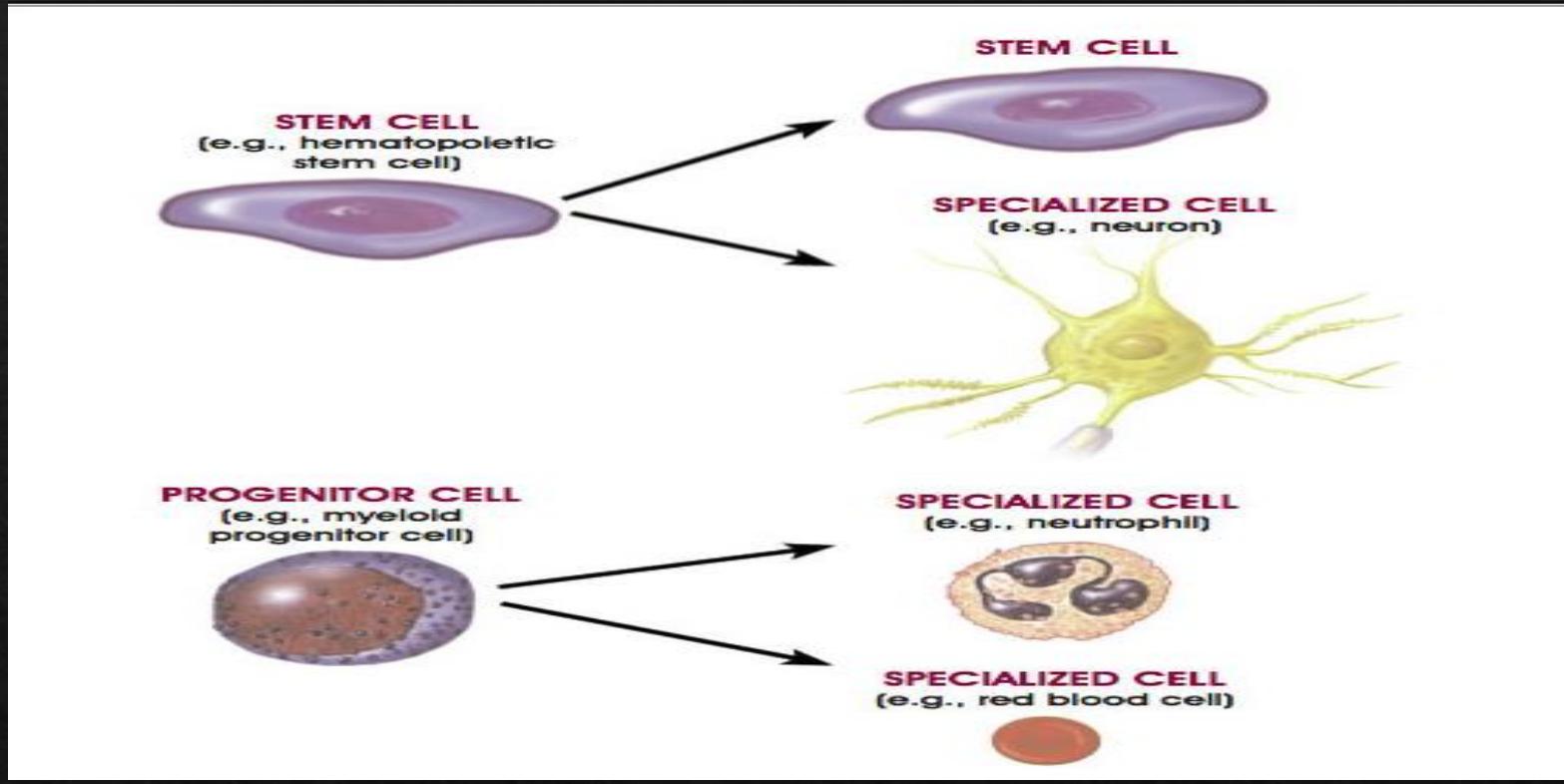
A quantitative system
for discriminating
induced pluripotent
stem cells, embryonic
stem cells

Developed by Mert
Tanyur ,Ipek Evren

Distinguishing stem cell types by methylation rate



Key words: biomarker, artificial neural network, support vector machines, induced pluripotent stem cells, embryonic stem cells, somatic cells, mathematical models



Abstract: Embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs) derived from somatic cells (SCs) provide promising resources for regenerative medicine and medical research, leading to a daily identification of new cell lines. However, an efficient system to discriminate the cell lines is lacking. Here, we developed a quantitative system to discriminate the three cell types, iPSCs, ESCs and SCs. The system contains DNA-methylation biomarkers and mathematical models ,including PCA.

Old but not gold

Traditionally, biomarkers derived from well-characterized individual molecules have been used to distinguish somatic cells (SCs) versus pluripotent cells (PCs), including iPSCs and ESCs [6,7]. PCR and immunostaining can be used to further aid the biomarkers in distinguishing SCs from PCs [6]. However, applying the biomarkers to inherent multipotent cell lines could mislead the results due to the instabilities of multipotent cell lines that vary with conditions [7]. For examples, the OCT4 biomarker, which was once thought to be an excellent marker for discriminating ESCs and SCs, is only transitionally expressed in ESCs and is not consistently expressed in different ESCs, especially in old ESCs [7]. Any single biomarker that was selected from a very limited number of samples is unlikely to be robust enough to classify novel stem cells when applied alone across various conditions [7]. In addition, most of the current biomarkers based on antibodies will fail to detect the protein signals that are of low abundance, and thus the antibody-based biomarkers naturally exhibit low sensitivity. The strategy developed for distinguishing SCs from PCs also may not work for discriminating ESCs and iPSCs due to their similarity.

Statictics rocks!

Biomarkers are selected unbiasly with PCA analysis. With 30 biomarkers ,or even with as few as 3 top biomarkers this system can discriminate ESCs from iPSCs with almost %100 percent accuracy.

ncbi.nlm.nih.gov

WisdomEra - Tibbi... Microsoft Office Ho... Customers Easy to use Online... WhatsApp Other book

NCBI Resources How To

All Databases

All Databases Assembly Biocollections BioProject BioSample BioSystems Books ClinVar Conserved Domains dbGaP dbVar Gene Genome GEO DataSets **GEO DataSets** GEO Profiles GTR HomoloGene Identical Protein Groups MedGen MeSH

COVID-19 is an emerging, rapidly evolving situation. Get public health information from CDC: <https://www.coronavirus.gov>. Get latest research from NIH: <https://www.nih.gov/coronavirus>.

NCBI

Center for Biotechnology Information advances science and improving access to biomedical and genomic information.

BLI | Mission | Organization | NCBI News & Blog

Download Transfer NCBI data to your computer

Learn Find help documents, attend a class or watch a tutorial

Analyze Identify an NCBI tool for your data analysis task

Research Explore NCBI research and collaborative projects

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

Methylations

Methylation microarrays were downloaded from GEO databases.

Data pre-
processed by
python.

ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL8490

WisdomEra - Tibbi... Microsoft Office Ho... Customers Easy to use Online... WhatsApp

NCBI

GEO Gene Expression Omnibus

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.coronavirus>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.

HOME SEARCH SITE MAP GEO Pub

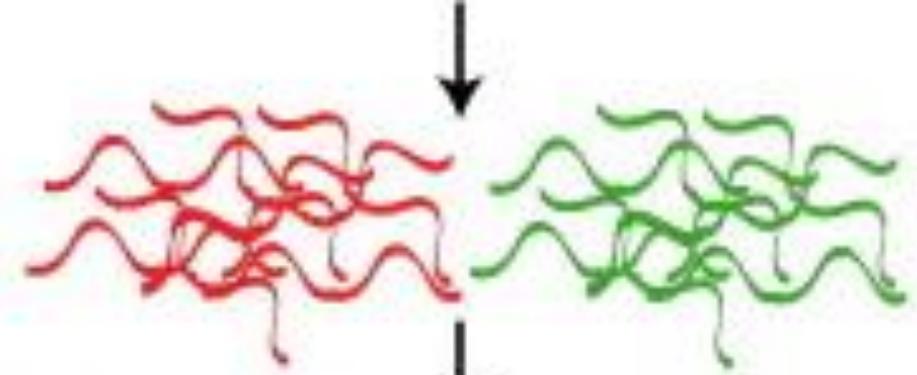
NCBI > GEO > Accession Display ?

Scope: Self Format: HTML Amount: Quick GEO accession: GPL8490 GO

Platform GPL8490 Query DataSets for GPL8490

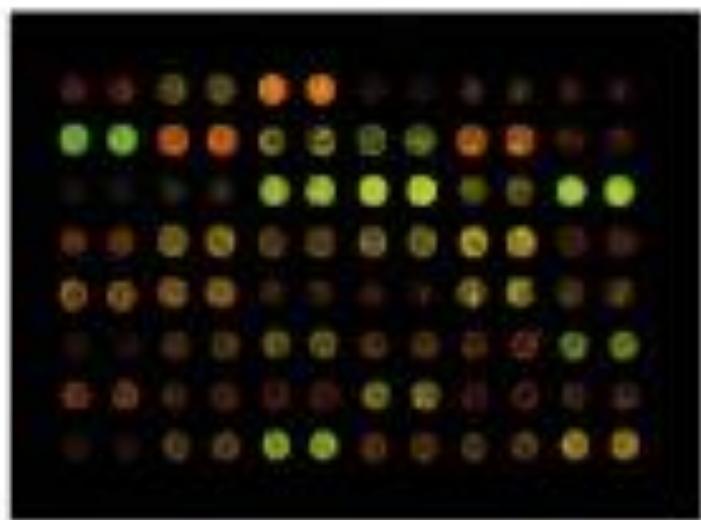
Status	Public on Apr 27, 2009		
Title	Illumina	HumanMethylation27	BeadChip
	(HumanMethylation27_270596_v.1.2)		
Technology type	oligonucleotide beads		
Distribution	commercial		
Organism	Homo sapiens		
Manufacturer	Illumina, Inc.		
Manufacture protocol	See manufacturer's website		
Description	HumanMethylation27 DNA Analysis BeadChip allows researchers to interrogate 27,578 highly informative CpG sites per sample at single-nucleotide resolution. This 12-sample BeadChip features content derived from the well-annotated NCBI CCDS database (Genome Build 36) and is supplemented with more than 1,000 cancer-related genes described in published literature. Probe content has been enriched to deeply cover more than 150 well-established cancer genes known to show differential		

We used
human
methylation
datasets



IV. Prepare CpG-island microarrays

V. Apply labeled samples to CpG-island array

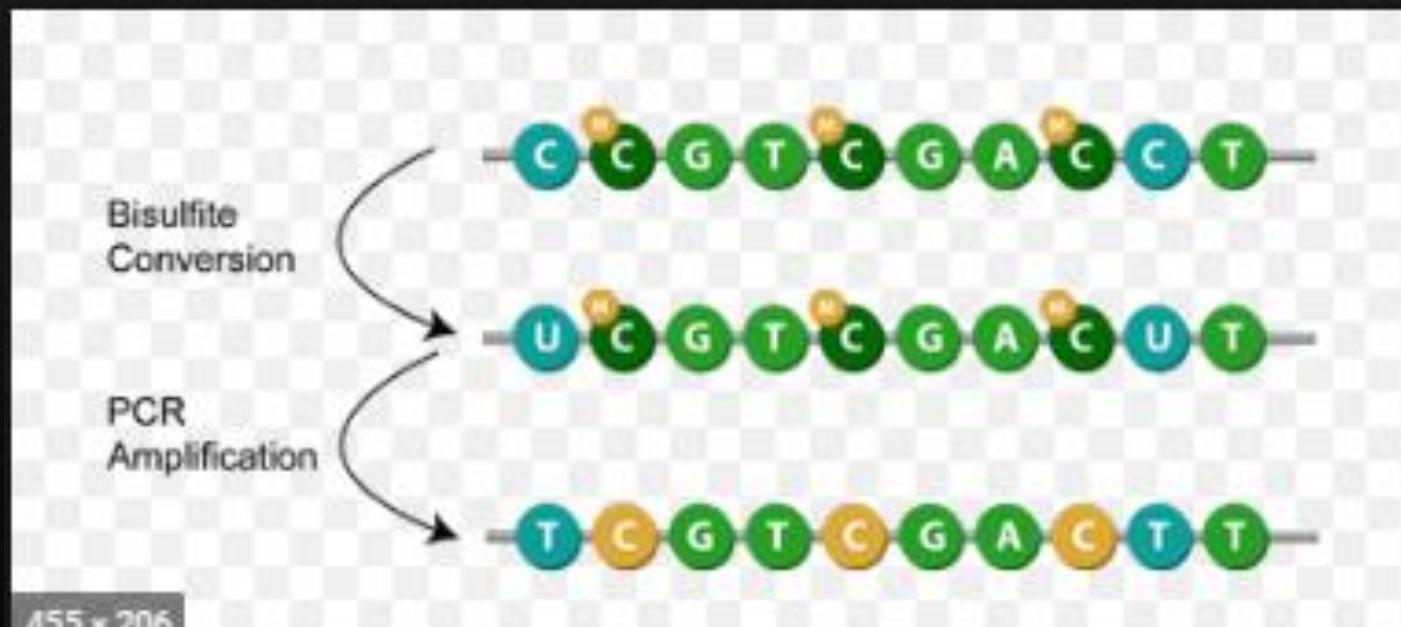


VI. Data analysis and confirmation

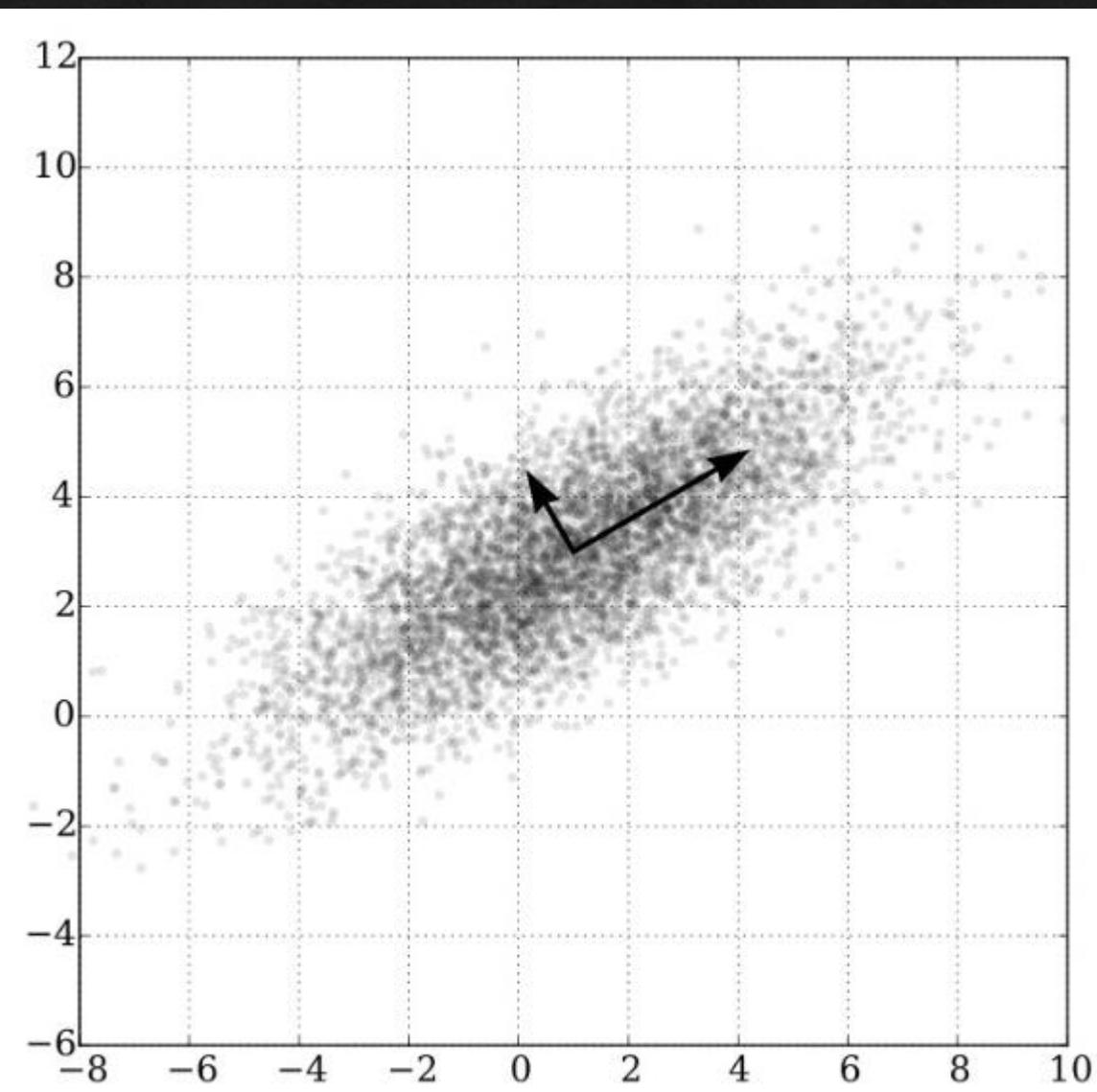
Lab-Method

- ❖ Methylation values measured as beta value ranging from 0
- ❖ CpG islands indicate the methylation regions.
- ❖ Two different colored beam is used to obtain a methylation intensity.

Bisulfite conversion is used to identify methylated bases.



How pca graphics look like



Some Statistical Terms

- ❖ Principal components analysis (PCA): PCA seeks a linear combination of variables such that the maximum variance is extracted from the variables. It then removes this variance and seeks a second linear combination which explains the maximum proportion of the remaining variance, and so on. This is called the principal axis method and results in orthogonal (uncorrelated) factors. PCA analyzes total (common and unique) variance

- ❖ Dimensionality : It is the number of random variables in a dataset or simply the number of features, or rather more simply, the number of columns present in your dataset.
- ❖ Correlation It shows how strongly two variable are related to each other. The value of the same ranges for -1 to +1. Positive indicates that when one variable increases, the other increases as well, while negative indicates the other decreases on increasing the former. And the modulus value of indicates the strength of relation.

- ❖ Orthogonal: Uncorrelated to each other, i.e., correlation between any pair of variables is 0.
- ❖ They represent the directions in which the data has maximum variance and also the directions in which the data is most spread out. If we are given a large dataset with multiple features, in which it would be difficult to select which of the variables (features) are the most important in determining the target, this PCA plays a huge role.

Why pca is great ?

- ❖ PCA reduces attribute space from a larger number of variables to a smaller number of factors and as such is a "non-dependent" procedure (that is, it does not assume a dependent variable is specified).• PCA is a dimensionality reduction or data compression method. The goal is dimension reduction and there is no guarantee that the dimensions are interpretable • To select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component.

- ❖ Thus, PCA is a method that brings together:
 - A measure of how each variable is associated with one another. (Covariance matrix.)
 - The directions in which our data are dispersed. (Eigenvectors.)
 - The relative importance of these different directions. (Eigenvalues.)PCA combines our predictors and allows us to drop the Eigenvectors that are relatively unimportant.

referances

<https://arxiv.org/ftp/arxiv/papers/1210/1210.5779.pdf>

I applied the same
methodology with this
study.

4. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, et al. (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131: 861-872.
5. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, et al. (2007) Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318: 1917-1920.
6. Carpenter MK, Rosler E, Rao MS (2003) Characterization and differentiation of human embryonic stem cells. *Cloning Stem Cells* 5: 79-88.
7. Goldman B (2008) Magic marker myths. *Nature Reports Stem Cells* doi:10.1038/stemcells.2008.26.
8. Muller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, et al. (2008) Regulatory networks define phenotypic classes of human stem cell lines. *Nature* 455: 401-405.
9. Muller FJ, Schuldert BM, Williams R, Mason D, Altun G, et al. (2011) A bioinformatic assay for pluripotency in human cells. *Nat Methods* 8: 315-317.
10. Newman AM, Cooper JB (2010) Lab-specific gene expression signatures in pluripotent stem cells. *Cell Stem Cell* 7: 258-262.
11. Guenther MG, Frampton GM, Soldner F, Hockemeyer D, Mitalipova M, et al. (2010) Chromatin structure and gene expression programs of human embryonic and induced pluripotent stem cells. *Cell Stem Cell* 7: 249-257.
12. Wang A, Huang K, Shen Y, Xue Z, Cai C, et al. (2011) Functional modules distinguish human induced pluripotent stem cells from embryonic stem cells. *Stem Cells Dev* 20: 1937-1950.
13. Goldstein DR GD, and Conlon EM (2002) Statistical issues in the clustering of gene expression data. *Statist Sinica* 12: 219-240.
14. Kim K, Doi A, Wen B, Ng K, Zhao R, et al. (2010) Epigenetic memory in induced pluripotent stem cells. *Nature* 467: 285-290.
15. Polo JM, Liu S, Figueredo ME, Kulalert W, Eminli S, et al. (2010) Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nat Biotechnol* 28: 848-855.
16. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, et al. (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* 471: 68-73.
17. Nazor KL, Altun G, Lynch C, Tran H, Harness JV, et al. (2012) Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. *Cell Stem Cell* 10: 620-634.
18. Lancashire LJ, Lemetre C, Ball GR (2009) An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies. *Brief Bioinform* 10: 315-329.
19. Oshima Y, Shinzawa H, Takenaka T, Furihata C, Sato H (2010) Discrimination analysis of human lung cancer cells associated with histological type and malignancy using Raman spectroscopy. *J Biomed Opt* 15: 017009.
20. Han M, Dai J, Zhang Y, Lin Q, Jiang M, et al. (2012) Support vector machines coupled with proteomics approaches for detecting biomarkers predicting chemotherapy resistance in small cell lung cancer. *Oncol Rep*.
21. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462: 315-322.
22. Bork S, Pfister S, Witt H, Horn P, Korn B, et al. (2010) DNA methylation pattern changes upon long-term culture and aging of human mesenchymal stromal cells. *Aging Cell* 9: 54-63.
23. Doi A, Park IH, Wen B, Murakami P, Aryee MJ, et al. (2009) Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet* 41: 1350-1353.
24. Chin MH, Mason MJ, Xie W, Volinia S, Singer M, et al. (2009) Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* 5: 111-123.

