Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS
ISSN 2454-6348

Original Research Article

DOI: 10.26479/2018.0404.09

DETERMINING PLURIPOTENCY OF A CELL BY AN *IN SILICO* METHOD

Abhishek Saini[1], Jai Gopal Sharma[1], Vimal Kishor Singh[2*]

1. Department of Biotechnology, Delhi Technological University, Delhi, India.

2. Tissue Engineering and Regenerative Medicine Laboratory, Department of Biomedical Engineering, Amity School of Engineering and Technology, Amity University Haryana, India.

ABSTRACT: While working with stem cells one of the questions that bother the researcher is about their potency to reprogram. So, if the pluripotency of cells is known beforehand, it would be an added advantage. Bioinformatics approaches have a viable solution to such concerns saving a lot on time and energy. Stem cell isolation and utilizing them for regeneration is itself a very sophisticated and dynamic concept. Thus working with such cells statistical and probabilistic aspects would be in consideration to tackle the active state. Induced pluripotent stem cells were first described by Shinya Yamanaka[1], and since then the approach towards the regenerative field of medicine has seen tremendous changes. By utilizing the bioinformatics tools and resources, few research groups have demonstrated the level of pluripotency of the cells. All available methods have different approaches and have their own set of limitations. After thorough research on such tools, we have designed a novel tool for determining the pluripotency. Unlike existing methods, our method doesn’t rely on particular cell type or specific file format to start with. It is a straightforward, quick and reliable way to check any cell’s pluripotency. The information regarding the gene expression on the pluripotent stem cells has been gathered at one place and compared with test cell, all with the help of programming languages viz. R, JAVA. Since the identification of cell is based on their genetic expression, the output is robust and reliable. It would be a key player in the field of regenerative medicine where pluripotent stem cells are being utilized for drug development and gene therapy. Further, for gene therapy and ex-vivo generation of cells determining the pluripotency of the source cells would not only increase the efficiency of the overall procedure but reduce the time and capital investment.

KEYWORDS: iPSCs, ESCs, POU5F1, NANOG, Pluripotency.

Corresponding Author: Dr. Vimal Kishor Singh* Ph.D.
Tissue Engineering and Regenerative Medicine Laboratory, Department of Biomedical
Engineering, Amity School of Engineering and Technology (ASET), Amity University Haryana,
Amity Education Valley, Gurgaon (Manesar), Haryana,India.
Email Address: vksingh@ggn.amity.edu

1. INTRODUCTION

Stem Cells are most promising in the regenerative field of medicine. With so many stem cells based therapies and protocols for ex-vivo generation of cells being developed throughout the globe. Several obstacles are faced by the researchers, and one of them is the determination of the pluripotency of any cell. As per definition, stem cells can be divided based on their potency viz. unipotent, multipotent, pluripotent and totipotent. A unipotent cell can only give rise to a single type of cell and self-renew, e.g., late erythroid precursor cells can give rise to only erythrocytes. A multipotent stem cell has a potential to differentiate into multiple cell types under given conditions, e.g., a Hematopoietic stem cell can differentiate to form red blood cells, platelets, macrophages or lymphocytes. Pluripotent stem cells have the potential to differentiate into any cells type of body. On the other hand, the totipotent stem cells can give rise to the whole organism as they have the potential to give rise to extraembryonic cells. The cell differentiation capacity to any of three cell lineages, i.e., Ectoderm, endoderm, Mesoderm is dependent upon the potency of that particular cell [2]. Pluripotency in embryonic stem cells is regulated by strict regulation of transcription factors and the epigenetic modifications [3], [4], [5], [6]. Besides transcription factors, cell surface markers, pathway-specific markers, and lectins, peptides markers also play a major role in regulating the pluripotency level of the cell [7]. Here, in this study we have used the bioinformatics tools, software and databases to overcome the situation of determining the pluripotency of any cell specially IPSCs. Initially, we have compared the already existing tools for the similar job, after analyzing the online tools we narrowed down on three commonly used tools viz. PluriTest, CellNet, and TertaScore. PluriTest is a DNA microarray-based tool in which researchers at Scripps have created a microarray database containing the genes expressing in embryonic stem cell and induced pluripotent stem cells. For determining the pluripotency of the given cell or cell line microarray file in *.idat (Raw intensity file) format, has to be created and uploaded on the website http://www.pluritest.org. Microsoft Silverlight is required to access the website, and the uploaded data is compared to the PluriTest microarray database. The results displayed show the information about pluripotency and probability of abnormalities if any in the stem cells [8]. Other two platforms uses Affymetrix generated .cel* file format. Cell Net is a network –biology based platform and claims to access the fidelity of cellular engineering more accurately. CellNet is available at (http://cellnet.hms.harvard.edu/) where .cel* format can be uploaded for analysis. A standalone version of the same is also available [9]. For characterization of pluripotent stem cells, teratoma formation is considered to be the standard gold assay. Teratoscore utilizes this property of pluripotent stem cells and distinguishes pluripotent stem cell derived Teratomas from malignant tumors and translates cell potency into a quantitative meas ure (http://benvenisty.huji.ac.il/teratoscore.php). It uses the gene expression data from all germ layers and extraembryonic tissues and represents the results in a scorecard form [10]. Therefore we begin our quest for a universal input file format, and after thorough research, we developed an input file in .txt* Text file format. The text file format is universally accepted and can easily be found for microarray analysis along with particular default formats. It also removes the dependence on specific file format like in case of other online tools. In-vitro identification and characterization of pluripotent stem cells include certain markers viz. cell surface markers, transcription factors, signal pathway related intracellular markers, enzymatic markers. Gene expression and genome-wide H3K4me3 and H3K27me3 expression are found to be much similar between ESCs and iPS cells. The reprogrammed iPSCs are found to be remarkably similar to naturally isolated pluripotent ESCs. In the following respects, thus confirming the identity, authenticity, and pluripotency of iPSCs to naturally isolated pluripotent stem cells, we came across some properties, which are commonly present in any IPSCs, and ESCs Telomeres, have ribonucleoprotein heterochromatin-like structure which is present at the ends of a chromosome that protect them from degradation as well as from being, detected as double-strand DNA breaks. Telomeres present in mammals is consist of tandem repeats of TTAGGG seq. and are involved in maintaining the sustainability of cell division, which is unrestricted by the maximum limit of ~50 cell divisions. Human ESCs shows high telomerase activity to maintain self-renewal and proliferation like properties. Similarly, IPSCs also shows high telomerase activity, and they show the expression of hTERT (human telomerase reverse transcriptase), which is a necessary component in the telomerase protein complex. IPSCs form Teratomas readily after nine weeks of their injection into immunodeficient mice. Teratomas are tumors consist of multiple lineages containing tissues, which are derived from the three germ layers; this is unlike other tumors, which typically are of only one cell type. Teratoma formation is also proving as a landmark test for the detection of pluripotency. Histones are compacting proteins that are structurally localized to DNA sequences that can affect their activity through various chromatin-related modifications. H3 histones which are associated with Nanog, Oct-3/4, and Sox2 were demethylated, which indicates that the expression of Oct-3/4, Nanog, and Sox2 [11], [12], [13], [14], [15], [16], [17].

2. MATERIALS AND METHODS

There is a total of 175 genes known as marker genes present in any pluripotent cell [7]. To validate this conclusion, an interaction network of these marker genes was generated using STRING, and after filtering the results by score provided by string, top scorer genes were got fetched out. Next step was to create a method which can identify pluripotency level of any cell. So, for this microarray data files (Control sample) were collected from NCBI's GEO (Gene Expression Omnibus) and *analyzed the results* using GEO2R (a GEO tool). Then, gene expression data file for that particular dataset was downloaded. After the preprocessing step which includes Gene matching (Using JAVA) and data arrangement; quantile function was adopted to calculate the threshold for any cell to be pluripotent. This generated threshold is based upon the Log FC value (Fold Change value) of the particular gene. After getting threshold, another data for test sample was taken and again after repeating the same process (as done for control sample), after this the log FC value of test sample was checked. Now, our main concern is to check that at what level the cell contains pluripotency? Since the cells that pass the threshold could either be multipotent or totipotent, therefore, we have developed a JAVA program. This JAVA program is trained by giving it a particular range for particular key regulator gene in both of ESC and iPSC condition. The result of the program is divided into three categories viz. Highly Pluripotent, Partial pluripotent, Low Pluripotent. This approach is a novel work for pluripotency determination as it uses the text format input file and a new robust method has been developed using Bioinformatics as a tool [18], [19], [20], [21], [22], [23], [24].

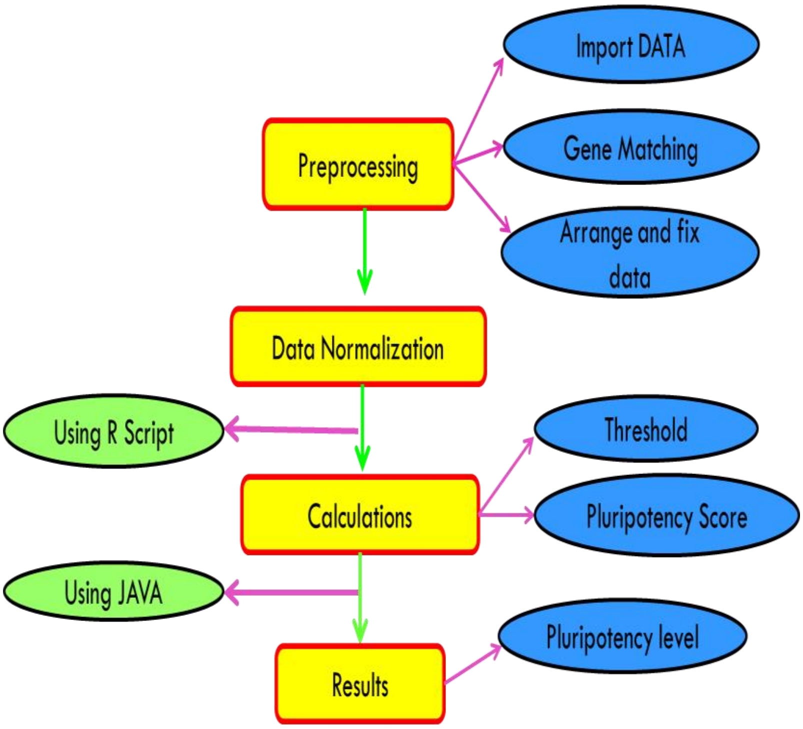


Fig.1: Defining various stages used in this method for finding pluripotency of any cell.

3. RESULTS AND DISCUSSION

The list of marker genes present in any stem cell which decides the pluripotency level of an undifferentiated cell was created. These genes were isolated by using different techniques viz. MACS (Magnetic Cell Sorting Technique) and FCM (flow cytometry) technique which is one of the most effective cell isolating methods [7].

Table 1: List of 175 signature markers used for identifying Stem cells.

SUV39H1	SMAD1	LECTINS	KLF4	CD57	CD86	STELLA
SUV39H2	SMAD5	CD133	NANOG	CD58	CD87	TRA-2-54
EHMT2	SMAD8	CD96	REX1	CD59	CD88	CD45
EHMT1	SMAD4	CD34	UTF1	CD60	CD89	CD56
SETDB1	SMAD2	CD38	ZFX	CD61	CD90	CD85
RING1B	SMAD3	CD45	TBN	CD62	CD326	ECSA
EZH2	BETA CATENIN	CD46	FOXD3	CD63	CD9	TM4SF
EED	SSEA1	CD47	HMGA2	CD64	CD55	TRA-2-49
SUZ12	CD15	CD48	NAC1	CD65	CD59	OCT4
DICER1	SSEA3	CD49	GCNF	CD66	CD24	SOX2
DNMT1	SSEA4	CD50	NR6A1	CD67	CD44	CD54
DNMT3a	CD324	CD51	STAT3	CD68	SATA3	CD55
DNMT3b	CD90	DRAP27	LEF1	CD69	NCA1	CD83
DNMT3L	CD117	P24	TCF3	CD70	ALDH1	CD84
CXXC1	CD326	CKIT	SALL4	CD71	MUSASHI-1	DPPA3
BRG1	CD9	SCFR	FBXO15	CD72	LgR5	CD82
SMARCA4	CD29	THY-1	ECAT11	CD73	PSCA	CD53
SMARCA5	CD24	TRA-1-60	FLJ10884	CD74	DCAMKL-1	OCT3
SMARCB1	CD59	TRA-1-81	L1TD1	CD75	TIM3	MRP1
SMARCC1	CD133	FRIZZLED5	ECAT1	CD76	BRCA1	DPPA2
MBD3	CD32	SCF	ECAT9	CD77	SDF1	
HIR A	CD49F	C-KIT	GDF3	CD78	CXCR4	
DPPA5	CD96	TDGF-1	TGF Beta	CD79	PSCA	
ESG1	HAS	CRIPTO	TCF1	CD80	CD96	
DPPA4	PROTECTIN	POU5F1	CD52	CD81	CD44	

To validate our findings of marker genes, we create an interaction network of all these above genes to check whether they are interaction partners or not. This result assures us that the given marker genes are interacting partners of key regulatory genes, i.e., OCT-4, NANOG, SOX2, KLF4 and are showing great interaction score with each other, which confirms us the presence of all these genes in the pluripotent cell. STRING provided us with the nodes which are commonly interacting with most of the genes and are having highest scores, i.e., POU5F1, NANOG, KLF4, SOX2, SALL4, SMAD2, SMAD4, and DPPA4. Another network was plotted using the resultant common genes which we got from the parent network.

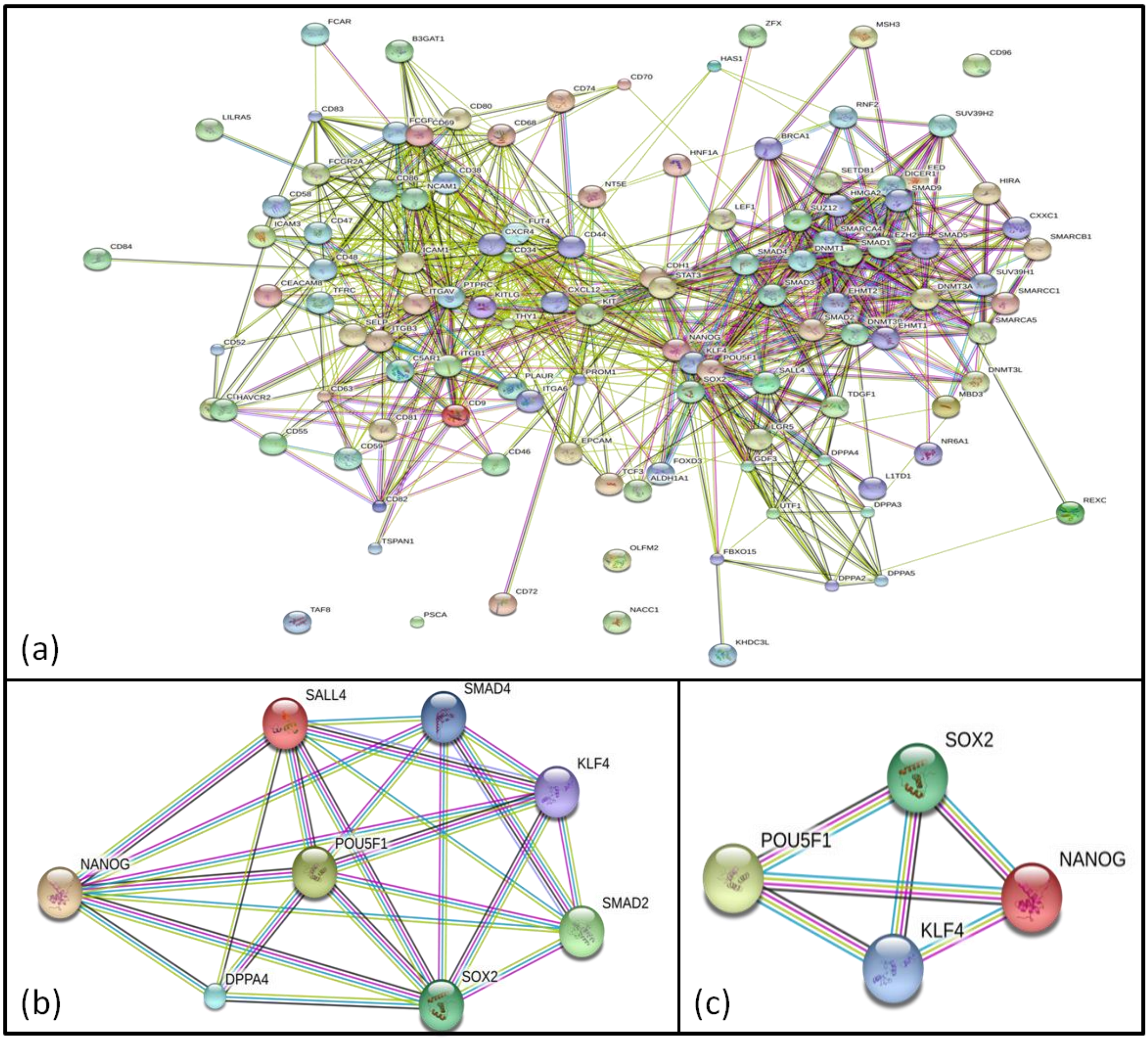
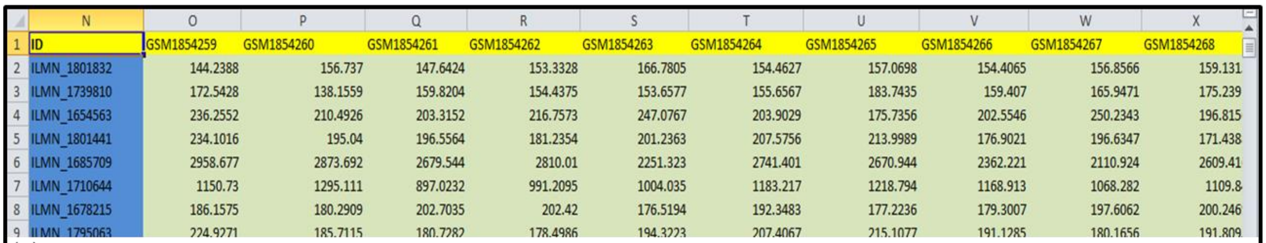


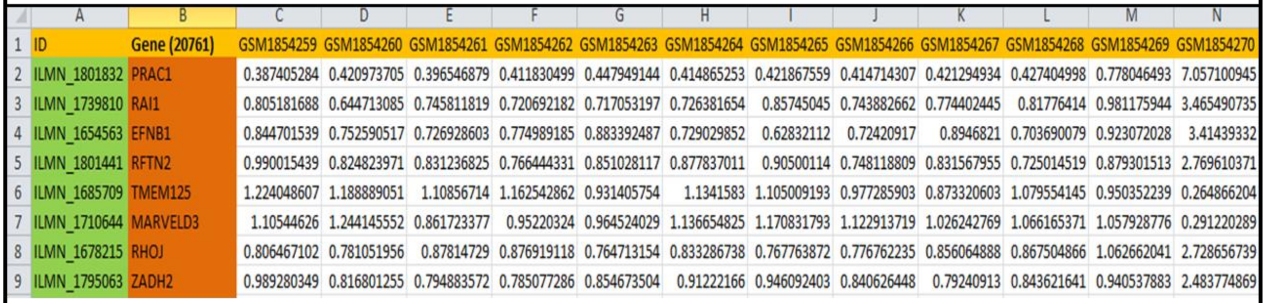
Fig.2: (a) Representation of Interaction network between various 175 Pluripotency Marker Genes identified by using STRING, each sphere (node) represents the particular gene. (b) A more specific STRING network is showing important genes responsible for pluripotency. (c) The final network with four key factors namely SOX2, NANOG, POU5F1, and KLF4.

We found that POU5F1, SOX2, NANOG, KLF4 are most commonly interacts. Thus an interaction network using STRING consisting of these four genes was plotted. Finally, genes with the most number of interactions and with highest no. of scores were obtained. By this we inference that the presence of these three genes and their expression values in microarray data could be the defining factor for the level of pluripotency in any cell. Now, our next step is to initially find out the state of any cell, i.e., whether it is pluripotent or not. For this, our first step is to determine the threshold for the genes of the cells. The cell has to pass this threshold before evaluating further. For this, we had taken Microarray data because this is the only way to check the expression value of the genes. The data we take is in TEXT file format, because of its availability. Firstly, the Microarray data containing Reference ID (Different for different types of analysis viz. “ILMN_xxxx” for Illumina analyzed data, “xxxx_s_at” for Affymetrix Data, “xxxxx” for Agilent Data), LogFC (Fold Change value showing Upregulation(+) and downregulation(-) of genes), Gene symbol etc. was downloaded for whole genome of human iPSC and ESC cell lines (GSE72078) from GEO datasets of National Centre for Biotechnology Information (NCBI) by analyzing with GEO2R (A GEO Tool available for visualization of Microarray data along with other relevant calculated statistical parameters). Now downloaded data was pasted into excel sheet with their respective gene expression values taken from series matrix file data. The data was then compared and matched with the expression values of available marker genes.



	N	O	P	Q	R	S	T	U	V	W	X
1	ID	GSM1854259	GSM1854260	GSM1854261	GSM1854262	GSM1854263	GSM1854264	GSM1854265	GSM1854266	GSM1854267	GSM1854268
2	ILMN_1801832	144.2388	156.737	147.6424	153.3328	166.7805	154.4627	157.0698	154.4065	156.8566	159.131
3	ILMN_1739810	172.5428	138.1559	159.8204	154.4375	153.6577	155.6567	183.7435	159.407	165.9471	175.239
4	ILMN_1654563	236.2552	210.4926	203.3152	216.7573	247.0767	203.9029	175.7356	202.5546	250.2343	196.815
5	ILMN_1801441	234.1016	195.04	196.5564	181.2354	201.2363	207.5756	213.9989	176.9021	196.6347	171.438
6	ILMN_1685709	2958.677	2873.692	2679.544	2810.01	2251.323	2741.401	2670.944	2362.221	2110.924	2609.41
7	ILMN_1710644	1150.73	1295.111	897.0232	991.2095	1004.035	1183.217	1218.794	1168.913	1068.282	1109.8
8	ILMN_1678215	186.1575	180.2909	202.7035	202.42	176.5194	192.3483	177.2236	179.3007	197.6062	200.246
9	ILMN_1795063	224.9271	185.7115	180.7282	178.4986	194.3223	207.4067	215.1077	191.1285	180.1656	191.809

(a)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	Gene (20761)	GSM1854259	GSM1854260	GSM1854261	GSM1854262	GSM1854263	GSM1854264	GSM1854265	GSM1854266	GSM1854267	GSM1854268	GSM1854269	GSM1854270
2	ILMN_1801832	PRAC1	0.387405284	0.420973705	0.396546879	0.411830499	0.447949144	0.414865253	0.421867559	0.414714307	0.421294934	0.427404998	0.778046493	7.057100945
3	ILMN_1739810	RAI1	0.805181688	0.644713085	0.745811819	0.720692182	0.717053197	0.726381654	0.85745045	0.743882662	0.774402445	0.81776414	0.981175944	3.465490735
4	ILMN_1654563	EFNB1	0.844701539	0.752590517	0.726928603	0.774989185	0.883392487	0.729029852	0.62832112	0.72420917	0.8946821	0.703690079	0.923072028	3.41439332
5	ILMN_1801441	RFTN2	0.990015439	0.824823971	0.831236825	0.766444331	0.851028117	0.877837011	0.90500114	0.748118809	0.831567955	0.725014519	0.879301513	2.769610371
6	ILMN_1685709	TMEM125	1.224048607	1.188889051	1.10856714	1.162542862	0.931405754	1.1341583	1.105009193	0.977285903	0.873320603	1.079554145	0.950352239	0.264866204
7	ILMN_1710644	MARVELD3	1.10544626	1.244145552	0.861723377	0.95220324	0.964524029	1.136654825	1.170831793	1.122913719	1.026242769	1.066165371	1.057928776	0.291220289
8	ILMN_1678215	RHOJ	0.806467102	0.781051956	0.87814729	0.876919118	0.764713154	0.833286738	0.767763872	0.776762235	0.856064888	0.867504866	1.062662041	2.728656739
9	ILMN_1795063	ZADH2	0.989280349	0.816801255	0.794883572	0.785077286	0.854673504	0.91222166	0.946092403	0.840626448	0.79240913	0.843621641	0.940537883	2.483774869

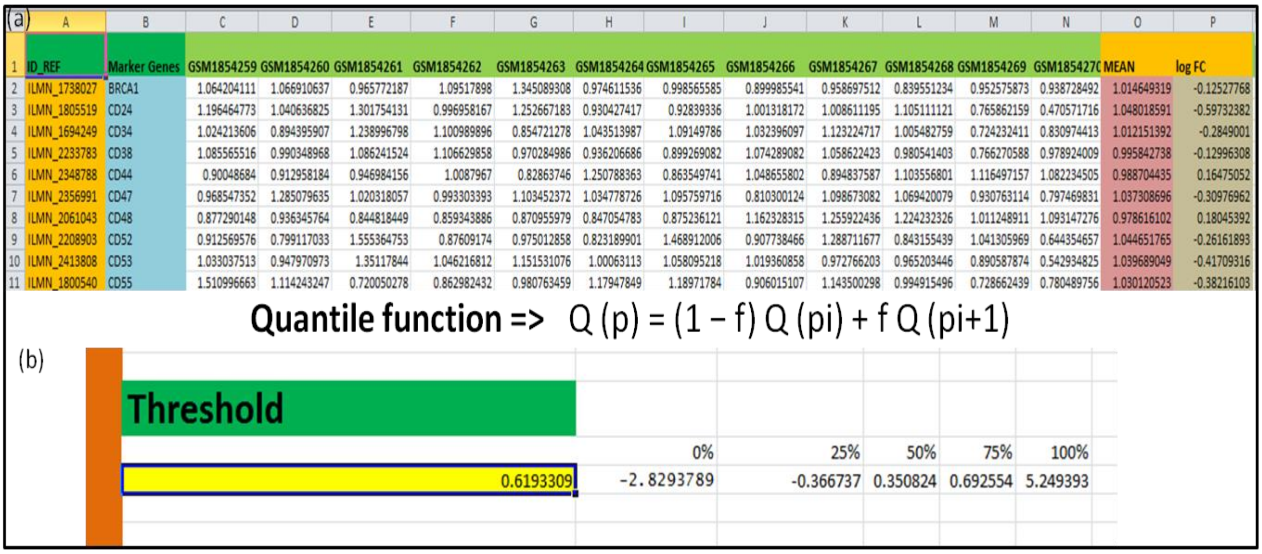
(b)

$$\bullet \text{Mean} = \sum (xi) / ni$$

$$\text{Normalization} = \frac{\text{(Expression value of Gene in particular cell line)}}{\text{(Mean of exp. Values of same gene for all cell lines present in a single microarray text file)}}$$

Fig.3 (a) Table is showing Non-Normalized Microarray data (GSE72078) in excel sheet collected from GEO datasets. (b) Normalized microarray data (using Simple normalization function).

After this, we matched our 175 Marker genes with the existing list of NCBI and downloaded expression dataset’s gene list by using JAVA program (specially designed for a finding of no. of marker genes present in the microarray data.



(a)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	ID_REF	Marker Genes	GSM1854259	GSM1854260	GSM1854261	GSM1854262	GSM1854263	GSM1854264	GSM1854265	GSM1854266	GSM1854267	GSM1854268	GSM1854269	GSM185427C	MEAN	log FC
2	ILMN_1738027	BRCA1	1.064204111	1.066910637	0.965772187	1.09517898	1.345089308	0.974611536	0.998565585	0.899985541	0.958697512	0.839551234	0.952575873	0.938728492	1.014649319	-0.12527768
3	ILMN_1805519	CD24	1.196464773	1.040636825	1.301754131	0.996958167	1.252667183	0.930427417	0.92839336	1.001318172	1.008611195	1.105111121	0.765862159	0.470571716	1.048018591	-0.59732382
4	ILMN_1694249	CD34	1.024213606	0.894395907	1.238996798	1.100989896	0.854721278	1.043513987	1.09149786	1.032396097	1.123224717	1.005482759	0.724232411	0.830974413	1.012151392	-0.2849001
5	ILMN_2233783	CD38	1.085565516	0.990348968	1.086241524	1.106629858	0.970284986	0.936206686	0.899269082	1.074289082	1.058622423	0.980541403	0.766270588	0.978924009	0.995842738	-0.12996308
6	ILMN_2348788	CD44	0.90048684	0.912958184	0.946984156	1.0087967	0.82863746	1.250788363	0.863549741	1.048655802	0.894837587	1.103556801	1.116497157	1.082234505	0.988704435	0.16475052
7	ILMN_2356991	CD47	0.968547352	1.285079635	1.020318057	0.993303393	1.103452372	1.034778726	1.095759716	0.810300124	1.098673082	1.069420079	0.930763114	0.797469831	1.037308696	-0.30976962
8	ILMN_2061043	CD48	0.877290148	0.936345764	0.844818449	0.859343886	0.870955979	0.847054783	0.875236121	1.162328315	1.255922436	1.224232326	1.011248911	1.093147276	0.978616102	0.18045392
9	ILMN_2208903	CD52	0.912569576	0.799117033	1.555364753	0.87609174	0.975012858	0.823189901	1.468912006	0.907738466	1.288711677	0.843155439	1.041305969	0.644354657	1.044651765	-0.26161893
10	ILMN_2413808	CD53	1.033037513	0.947970973	1.35117844	1.046216812	1.151531076	1.00063113	1.058095218	1.019360858	0.972766203	0.965203446	0.890587874	0.542934825	1.039689049	-0.41709316
11	ILMN_1800540	CD55	1.510996663	1.114243247	0.720050278	0.862982432	0.980763459	1.17947849	1.18971784	0.906015107	1.143500298	0.994915496	0.728662439	0.780489756	1.030120523	-0.38216103

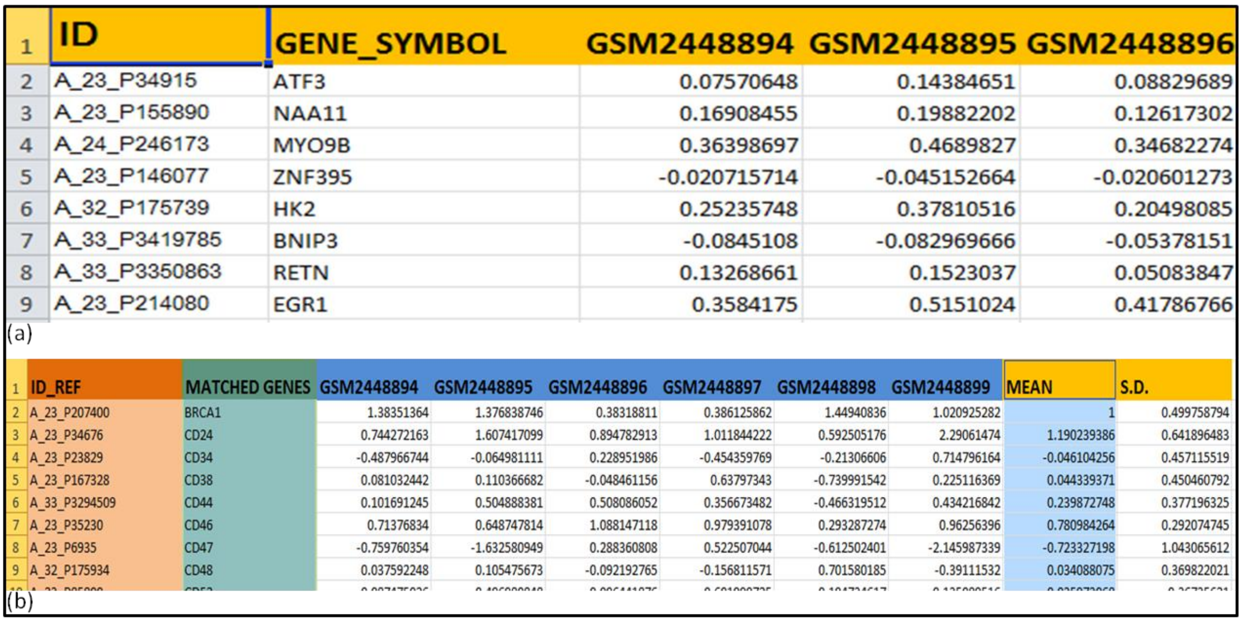
Quantile function => Q (p) = (1 – f) Q (pi) + f Q (pi+1)

(b)

Threshold	0%	25%	50%	75%	100%
0.6193309	-2.8293789	-0.366737	0.350824	0.692554	5.249393

Fig.4 (a) Table is showing marker genes with reference IDs and their respective expression values in each cell line. Also, Calculated Mean and log FC values are shown. (b) The calculated Threshold value for detection of pluripotency using Quantile Normalization method through R script (Threshold = 0.6193309).

Total 68 Marker Genes are matched with the Genes of Microarray Data of GSE72078 cell lines. After normalization we took the data for prediction of Threshold using Quantile function through R Studio, this approach gives five different quantile calculated values including min. And max.; after taking the mean for these values, we got our Threshold, i.e., 0.62, which implies that if any cell passes this threshold, then it could say to be pluripotent. Now, the next step is to check the pluripotency status for a sample dataset (Test Sample). For this, we had taken a colon IPSC Cell line. The microarray data was downloaded for Colon IPSC's (GSE93228- Cell lines iPSC CRL1831 (induced pluripotent stem cells) derived from normal colon CRL1831 cells in 3D cell culture conditions and subjected to ionizing radiation doses) and then after arranging and preprocessing the data we check whether the cell lines are Stem cell lines or not. If they all are stem cell lines then we simply check their pluripotency score by taking the quantile normalization of their Log FC value, but if the data consist of both differentiated and undifferentiated cell lines then we have to take mean of each cell line's expression values and then match with the Threshold limit to check whether they are passing the set Threshold value or not. If the resultant score is less than the threshold, then that cell could be either Unipotent, totipotent, multipotent or differentiated somatic cell line.



	ID	GENE_SYMBOL	GSM2448894	GSM2448895	GSM2448896
2	A_23_P34915	ATF3	0.07570648	0.14384651	0.08829689
3	A_23_P155890	NAA11	0.16908455	0.19882202	0.12617302
4	A_24_P246173	MYO9B	0.36398697	0.4689827	0.34682274
5	A_23_P146077	ZNF395	-0.020715714	-0.045152664	-0.020601273
6	A_32_P175739	HK2	0.25235748	0.37810516	0.20498085
7	A_33_P3419785	BNIP3	-0.0845108	-0.082969666	-0.05378151
8	A_33_P3350863	RETN	0.13268661	0.1523037	0.05083847
9	A_23_P214080	EGR1	0.3584175	0.5151024	0.41786766

(a)

	ID_REF	MATCHED GENES	GSM2448894	GSM2448895	GSM2448896	GSM2448897	GSM2448898	GSM2448899	MEAN	S.D.
2	A_23_P207400	BRCA1	1.38351364	1.376838746	0.38318811	0.386125862	1.44940836	1.020925282	1	0.499758794
3	A_23_P34676	CD24	0.744272163	1.607417099	0.894782913	1.011844222	0.592505176	2.29061474	1.190239386	0.641896483
4	A_23_P23829	CD34	-0.487966744	-0.064981111	0.228951986	-0.454359769	-0.21306606	0.714796164	-0.046104256	0.457115519
5	A_23_P167328	CD38	0.081032442	0.110366682	-0.048461156	0.63797343	-0.739991542	0.225116369	0.044339371	0.450460792
6	A_33_P3294509	CD44	0.101691245	0.504888381	0.508086052	0.356673482	-0.466319512	0.434216842	0.239872748	0.377196325
7	A_23_P35230	CD46	0.71376834	0.648747814	1.088147118	0.979391078	0.293287274	0.96256396	0.780984264	0.292074745
8	A_23_P6935	CD47	-0.759760354	-1.632580949	0.288360808	0.522507044	-0.612502401	-2.145987339	-0.723327198	1.043065612
9	A_32_P175934	CD48	0.037592248	0.105475673	-0.092192765	-0.156811571	0.701580185	-0.39111532	0.034088075	0.369822021

(b)

Fig.5: (a)Table Showing Test Sample consisting of GENE symbol with their expression values in respective cell lines of Colon IPSC (total 6 IPSC cell lines are taken after neglecting somatic cell lines data). (b)Table after matching our Marker genes with the Genes of Microarray Data of GSE93228 cell lines we obtained 71 matched entries by using JAVA developed program.

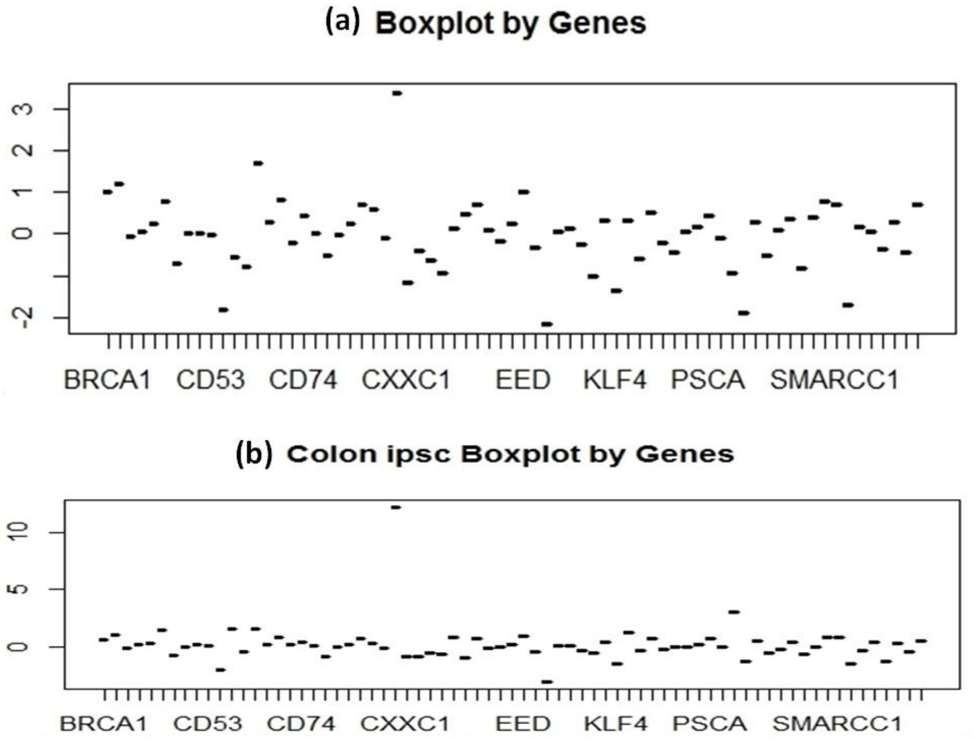
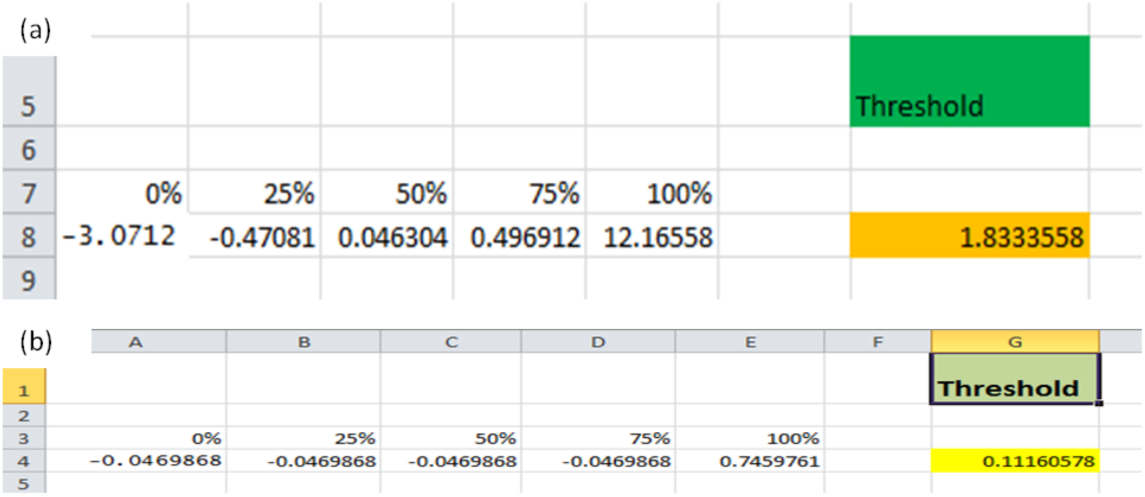


Fig.6: (a) Boxplot is showing non-normalized gene expression data of GSE93228, depicting discreteness of the values. (b) Boxplot showing Normalized gene expression values in a linear manner



(a)

5							Threshold
6							
7	0%	25%	50%	75%	100%		
8	-3.0712	-0.47081	0.046304	0.496912	12.16558		1.8333558
9							

(b)

	A	B	C	D	E	F	G
1							Threshold
2							
3	0%	25%	50%	75%	100%		
4	-0.0469868	-0.0469868	-0.0469868	-0.0469868	0.7459761		0.11160578
5							

Fig.7: (a) The Pluripotency score is calculated again using R Script as earlier, and it passed the Threshold value. Hence we can say that the Cell line for Microarray data of GSE 93228 sample is pluripotent. (b)After the calculation of pluripotency score using Quantile function we deliberately checked mast cell data which could not pass the Threshold value.

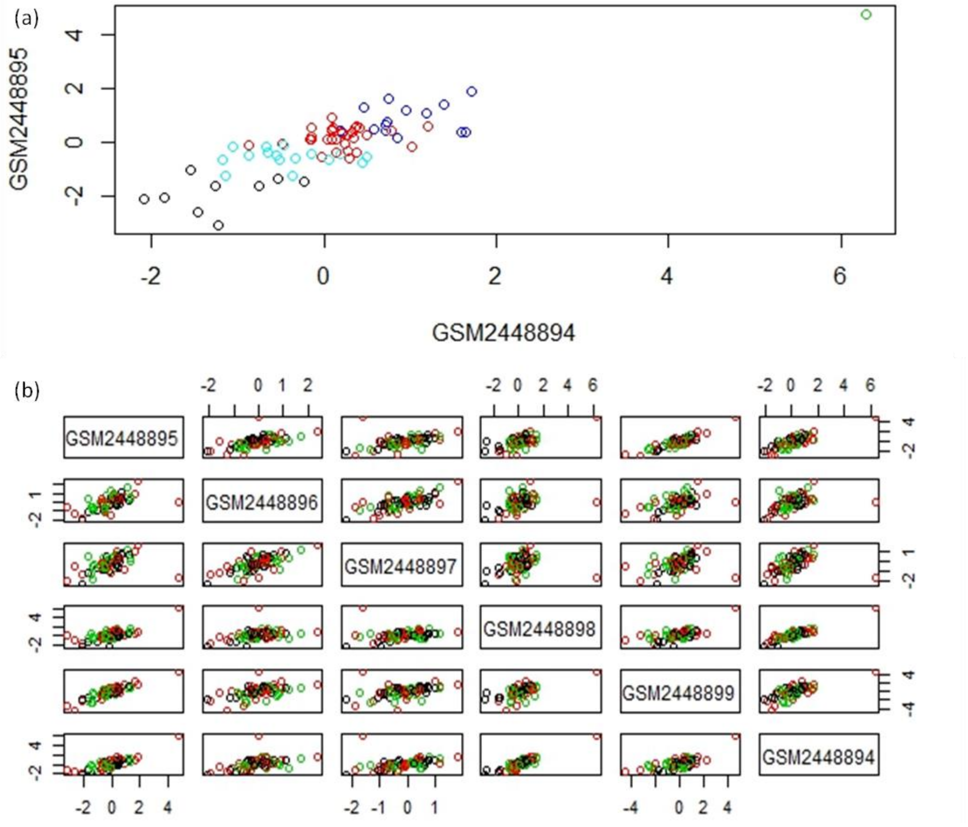


Fig.8: (a) Graph is showing clustered cell line from test sample by Gene expression data by using K-Means clustering through R Script for the development of Hierarchical clustering in Heatmap. (b)Figure showing K-Means clustered data of all six cell lines of test data sets created using R Script.

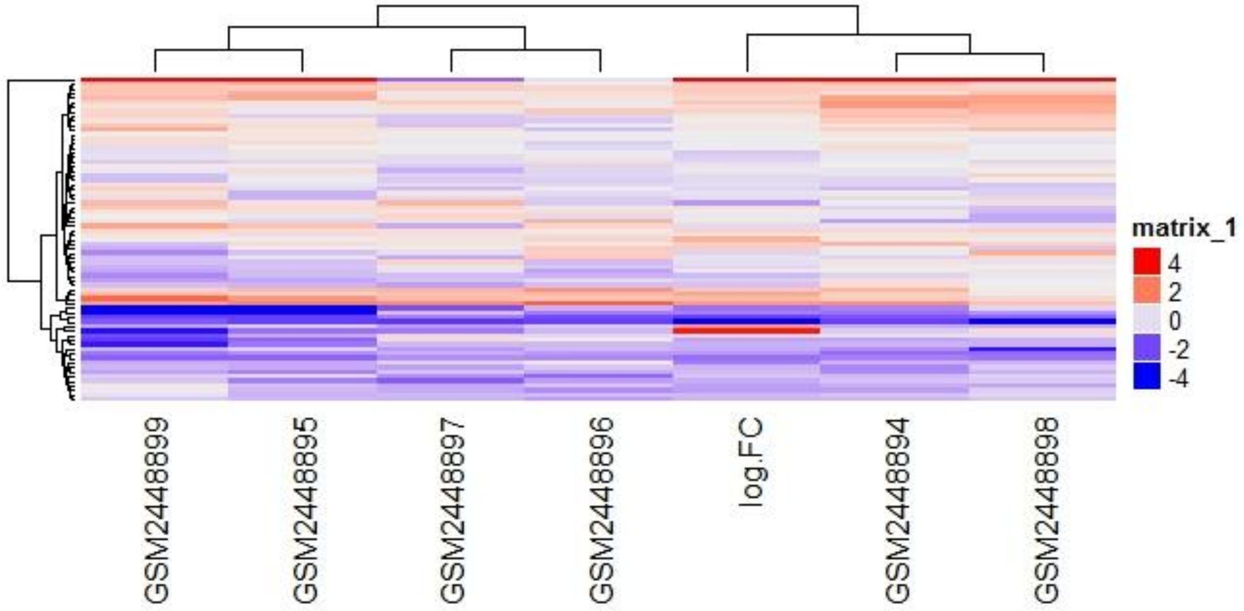


Fig.9: Heatmap Generated for all test sample's cell line's gene expression data using the R program.

After this, we analyze the level of Pluripotency using JAVA Program. For this, we initially set up a range parameter for all key regulator genes in three different ways viz.

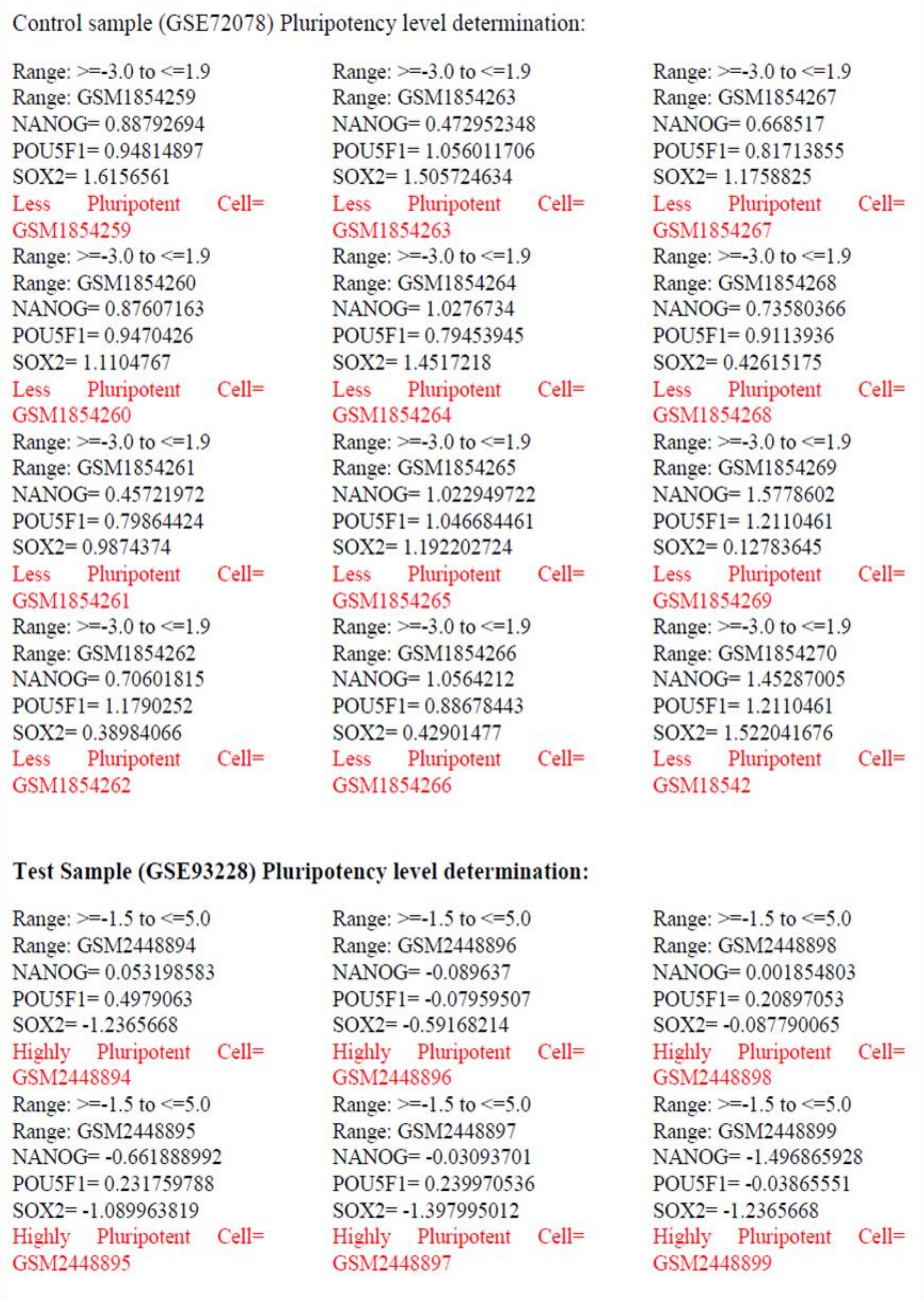
1) Highly Pluripotent cells.

2) Partially Pluripotent cells.

3) Less Pluripotent cells.

Table 2	IPSC Range		ESC Range	
	Is Data Manually Normalized or Not			
	YES	NO	YES	NO
1)	5.0 to 9.0	-1.5 to 5.0	2.0 to 3.0	-0.25 to 2.0
2)	2.0 to 4.9	-2.0 to -1.51	1.5 to 1.9	-3.0 to -0.249
3)	-3.0 to 1.9	-3.0 to -2.1	1.2 to 1.49	-15.0 to -2.9

This Range is based upon the manually compared and calculated expression values of key regulators from Different samples (Microarray Samples) viz. (GSE72078, GSE76282, GSE42445, etc.) present in GEO datasets and depend upon the condition that either the data is pre-normalized or manually normalized. For both, the conditions the range is individually provided in each case of ESC as well as IPSC.



Control sample (GSE72078) Pluripotency level determination:

```
Range: >=-3.0 to <=1.9
Range: GSM1854259
NANOG= 0.88792694
POU5F1= 0.94814897
SOX2= 1.6156561
Less    Pluripotent    Cell=
GSM1854259
Range: >=-3.0 to <=1.9
Range: GSM1854260
NANOG= 0.87607163
POU5F1= 0.9470426
SOX2= 1.1104767
Less    Pluripotent    Cell=
GSM1854260
Range: >=-3.0 to <=1.9
Range: GSM1854261
NANOG= 0.45721972
POU5F1= 0.79864424
SOX2= 0.9874374
Less    Pluripotent    Cell=
GSM1854261
Range: >=-3.0 to <=1.9
Range: GSM1854262
NANOG= 0.70601815
POU5F1= 1.1790252
SOX2= 0.38984066
Less    Pluripotent    Cell=
GSM1854262
```

```
Range: >=-3.0 to <=1.9
Range: GSM1854263
NANOG= 0.472952348
POU5F1= 1.056011706
SOX2= 1.505724634
Less    Pluripotent    Cell=
GSM1854263
Range: >=-3.0 to <=1.9
Range: GSM1854264
NANOG= 1.0276734
POU5F1= 0.79453945
SOX2= 1.4517218
Less    Pluripotent    Cell=
GSM1854264
Range: >=-3.0 to <=1.9
Range: GSM1854265
NANOG= 1.022949722
POU5F1= 1.046684461
SOX2= 1.192202724
Less    Pluripotent    Cell=
GSM1854265
Range: >=-3.0 to <=1.9
Range: GSM1854266
NANOG= 1.0564212
POU5F1= 0.88678443
SOX2= 0.42901477
Less    Pluripotent    Cell=
GSM1854266
```

```
Range: >=-3.0 to <=1.9
Range: GSM1854267
NANOG= 0.668517
POU5F1= 0.81713855
SOX2= 1.1758825
Less    Pluripotent    Cell=
GSM1854267
Range: >=-3.0 to <=1.9
Range: GSM1854268
NANOG= 0.73580366
POU5F1= 0.9113936
SOX2= 0.42615175
Less    Pluripotent    Cell=
GSM1854268
Range: >=-3.0 to <=1.9
Range: GSM1854269
NANOG= 1.5778602
POU5F1= 1.2110461
SOX2= 0.12783645
Less    Pluripotent    Cell=
GSM1854269
Range: >=-3.0 to <=1.9
Range: GSM1854270
NANOG= 1.45287005
POU5F1= 1.2110461
SOX2= 1.522041676
Less    Pluripotent    Cell=
GSM18542
```

Test Sample (GSE93228) Pluripotency level determination:

```
Range: >=-1.5 to <=5.0
Range: GSM2448894
NANOG= 0.053198583
POU5F1= 0.4979063
SOX2= -1.2365668
Highly    Pluripotent    Cell=
GSM2448894
Range: >=-1.5 to <=5.0
Range: GSM2448895
NANOG= -0.661888992
POU5F1= 0.231759788
SOX2= -1.089963819
Highly    Pluripotent    Cell=
GSM2448895
```

```
Range: >=-1.5 to <=5.0
Range: GSM2448896
NANOG= -0.089637
POU5F1= -0.07959507
SOX2= -0.59168214
Highly    Pluripotent    Cell=
GSM2448896
Range: >=-1.5 to <=5.0
Range: GSM2448897
NANOG= -0.03093701
POU5F1= 0.239970536
SOX2= -1.397995012
Highly    Pluripotent    Cell=
GSM2448897
```

```
Range: >=-1.5 to <=5.0
Range: GSM2448898
NANOG= 0.001854803
POU5F1= 0.20897053
SOX2= -0.087790065
Highly    Pluripotent    Cell=
GSM2448898
Range: >=-1.5 to <=5.0
Range: GSM2448899
NANOG= -1.496865928
POU5F1= -0.03865551
SOX2= -1.2365668
Highly    Pluripotent    Cell=
GSM2448899
```

Fig.10: Result for Pluripotency level determination through JAVA program

For Pluripotency level determination we first took the control sample. Our program first checks whether the cell lines are IPSC or ESC. Then after confirming that the cell lines are for IPSC. It asked for whether the microarray data was manually normalized or not and then according to our entries for IPSC cell line with manually normalized data, the program took to consider the range for this condition and gave us the results that in which category or level the given sample is lying. In our control results show that the cells are least pluripotent in normalized IPSC range. Same is the case of test sample our program initially took the same step as done for control and then decides the level of pluripotency. Here, in our test sample, the condition came out for non manually normalized IPSC data and hence, we got the results for that condition range. We also took test sample datasets from different other arrays too like GSE92706 and GSE73330 which were found to be passed and failed respectively. For confirming our results, we cross-check our results with PLURITEST by taking the above IDs data in (.idat*) raw intensity file format, and after analyzing with PLURITEST, the result we got are surprisingly as same as ours. By, this we conclude that the method which we develop to test pluripotency using (.txt) text file format is worth to work with and giving favorable as well as satisfactory results. Here we have shown only results for GSE92706. The comparable results of tested sample with different approaches (i.e., using a text format and .idat format) are shown in figures. From our study, we found that the potency determination is a key factor of gene expression analysis and by considering only the gene effect we could determine various activities of cells including pluripotency. This in vitro approach is competing with other pre available online tools. Such tools are still dealing with some bugs as by considering the only limited number of files and with limited file format reliability. Also, their dependency on online servers are making them not 100% fit for potency determination, but in our case, the approach is working on the text file based logic of creating and arranging raw text file into matched and arranged file format concerning the gene expression values and Log FC. The level of pluripotency which we are calculating through JAVA determines us three levels, where each level gives us an idea about the potency of that particular cell to be differentiated into the basic three lineages. Highest level determines the potency of differentiating into all three lineages, whereas partial and low level determines us the differentiation into either two or one of the three lineages respectively. This approach gave us the way for determining the potency using computer programming language JAVA and statistical method based R Script, which was used in arranging data according to the matched marker genes and finally in the determination of Level of pluripotency using various ranges. Several graphs and plots viz. boxplots, clustering graphs, and heatmap were also developed using an R script. This approach will surely provide open access for identifying pluripotency and understanding the working and expression nature of various genes involved in reprogramming strategies.

4. CONCLUSION

Microarray data proves beneficial in many regards. To get detailed info about any process related to protein or gene expression or their interaction we do need to take help from it. As in the above study, we found that to get pluripotency test of any cell sample. First, we have to access gene expression data of that particular cell. Followed by grabbing and arranging that data in a proper format, we will be able to fetch pluripotency data after checking whether the log FC value for that test sample either passing the threshold or not. Using this approach we are now in a state to tackle various problems associated with pre-existing online tools. We can make our tool based on this approach which will be free from various bugs that are present in existing tools; also we can identify which gene is devoting more in making any cell pluripotent. Through this approach, we can determine the pluripotency state of test cell regardless of any particular platform or any file format.

5. ACKNOWLEDGEMENT

We thank the honorable Vice Chancellor, Delhi Technological University, for providing essential support for carrying out the research.

6. CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, Tomoda K, Yamanaka S. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. cell. 2007 Nov 30;131(5):861-72.
2. Solter D. From teratocarcinomas to embryonic stem cells and beyond: a history of embryonic stem cell research. Nature Reviews Genetics. 2006 Apr;7(4):319.
3. Niwa H, Miyazaki JI, Smith AG. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nature genetics. 2000 Apr;24(4):372.
4. Mitsui K, Tokuzawa Y, Itoh H, Segawa K, Murakami M, Takahashi K, Maruyama M, Maeda M, Yamanaka S. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. cell. 2003 May 30;113(5):631-42.
5. Chambers I, Colby D, Robertson M, Nichols J, Lee S, Tweedie S, Smith A. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell. 2003 May 30;113(5):643-55.
6. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK. Core transcriptional regulatory circuitry in human embryonic stem cells. cell. 2005 Sep 23;122(6):947-56.
7. Zhao W, Ji X, Zhang F, Li L, Ma L. Embryonic stem cell markers. Molecules. 2012 May 25;17(6):6196-236.

8. Müller FJ, Schuldt BM, Williams R, Mason D, Altun G, Papapetrou EP, Danner S, Goldmann JE, Herbst A, Schmidt NO, Aldenhoff JB. A bioinformatic assay for pluripotency in human cells. Nature methods. 2011 Mar 6;8(4):315.
9. Cahan P, Li H, Morris SA, Da Rocha EL, Daley GQ, Collins JJ. CellNet: network biology applied to stem cell engineering. Cell. 2014 Aug 14;158(4):903-15.
10. Avior Y, Biancotti JC, Benvenisty N. TeratoScore: assessing the differentiation potential of human pluripotent stem cells by quantitative expression analysis of teratomas. Stem cell reports. 2015 Jun 9;4(6):967-74.
11. Gokhale PJ, Andrews PW. The development of pluripotent stem cells. Current opinion in genetics & development. 2012 Oct 31;22(5):403-8.
12. Niwa H. How is pluripotency determined and maintained?. Development. 2007 Feb 15;134(4):635-46.
13. Singh VK, Kalsan M, Kumar N, Saini A, Chandra R. Induced pluripotent stem cells: applications in regenerative medicine, disease modeling, and drug discovery. Frontiers in cell and developmental biology. 2015 Feb 2;3:2.
14. Liu Y, Cheng D, Li Z, Gao X, Wang H. The gene expression profiles of induced pluripotent stem cells (iPSCs) generated by a non-integrating method are more similar to embryonic stem cells than those of iPSCs generated by an integrating method. Genetics and molecular biology. 2012;35(3):693-700.
15. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, Wong KY. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nature genetics. 2006 Apr;38(4):431.
16. Perez-Iratxeta C, Andrade-Navarro MA, Wren JD. Evolving research trends in bioinformatics. Briefings in Bioinformatics. 2006 Oct 31;8(2):88-95.
17. Tiemann U, Marthaler AG, Adachi K, Wu G, Fischedick GU, Araúzo-Bravo MJ, Schöler HR, Tapia N. Counteracting activities of OCT4 and KLF4 during reprogramming to pluripotency. Stem cell reports. 2014 Mar 11;2(3):351-65.
18. Zhao JH, Tan Q. Integrated analysis of genetic data with R. Human genomics. 2006 Dec;2(4):258.
19. Nestor MW, Noggle SA. Standardization of human stem cell pluripotency using bioinformatics. Stem cell research & therapy. 2013 Jun;4(2):37.
20. Babu PB, Krishnamoorthy P. Applications of Bioinformatics Tools in Stem Cell Research: An Update. Journal of Pharmacy Research Vol. 2012 Sep;5(9):4863-6.

21. Dudoit S, Yang JY. Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. InThe analysis of gene expression data 2003 (pp. 73-101). Springer, New York, NY.

22. Som A, Harder C, Greber B, Siatkowski M, Paudel Y, Warsow G, Cap C, Schöler H, Fuellen G. The PluriNetWork: an electronic representation of the network underlying pluripotency in mouse, and its applications. PloS one. 2010 Dec 10;5(12):e15165.

23. Zhao W, Ji X, Zhang F, Li L, Ma L. Embryonic stem cell markers. Molecules. 2012 May 25;17(6):6196-236.

24. Huber W, Heydebreck AV, Vingron M. Analysis of microarray gene expression data. InHandbook of Statistical Genetics 2003 Jul. John Wiley & Sons, Ltd.