

“To develop an *In silico* method for the detection of pluripotency on the basis of comparative analysis of various Pluripotency determining tools”

A Major Project dissertation submitted in partial fulfilment of the requirement for the degree of

**Master of Technology
In
Bioinformatics**

Submitted by

RAZI KHAN

(2K15/BIO/11)

Delhi Technological University, Delhi, India

Under the supervision of

Dr. Vimal Kishor Singh



Department of Biotechnology

Delhi Technological University

Shahbad Daulatpur, Main Bawana Road,

Delhi – 110042, INDIA



CERTIFICATE

This is to certify that the M.Tech. Dissertation entitled “*To develop an In silico method for the detection of pluripotency on the basis of comparative analysis of various Pluripotency determining tools*”, submitted by RAZI KHAN (2K15/BIO/11) in partial fulfilment of the requirement for the award of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate’s own work carried out by him under my guidance.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honoring of any other degree.

Date:

Dr. Vimal Kishor Singh

(Project Mentor)

Department of Biotechnology

Delhi Technological University

Shahbad Bawana Road, Delhi-110042

DECLARATION

I, Razi Khan, hereby declare that the report entitled “*To develop an In silico method for the detection of pluripotency on the basis of comparative analysis of various Pluripotency determining tools*” submitted in partial fulfillment of the requirement for the award of the degree of Master of Technology, Delhi Technological University, is a record of original and independent research work done by me under the supervision and guidance of Dr. Vimal Kishor Singh, (O/I* Stem Cell Research Laboratory) Department of Biotechnology at Delhi Technological University, Delhi and the thesis has not formed the basis of the award of any Degree/Diploma/Associateship/Fellowship or other similar title to any candidate of any University/institution.

Date:

Signature of Candidate

ACKNOWLEDGEMENT

I would first like to thank my thesis advisor Dr.Vimal Kishor Singh, (O/I* Stem Cell Research Laboratory) Department of Biotechnology at Delhi Technological University. The door to Prof. Singh office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this thesis to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to express my sincere thanks to Prof. D. Kumar, H.O.D., Department of Biotechnology, D.T.U., Delhi for giving me this opportunity to undertake this project.

I also wish to express my gratitude to my teachers and other non-teaching staff members for providing us their valuable teachings with constant support and guidance.

I would also like to acknowledge Mr. Abhishek Saini (Ph.D. scholar) of the Stem cell research lab at DTU as the second reader of this thesis, who supported at every instance of my hurdles during the project and I am gratefully indebted for his very valuable comments on this thesis.

Also, my special thanks goes to my beloved friend Aniruddha (PG Student, Deptt. of Biotech., DTU) who helps me in understanding and seeking the programming part of my project and to all my buddies who are there with me for any kind of assistance.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Place : Delhi

RAZI KHAN

Date :

M.Tech. (Bioinformatics)

(2K15/BIO/11)

CONTENTS

TOPIC

Page No.

List of Figures

List of Tables

List of Abbreviations

1. ABSTRACT	9
2. INTRODUCTION	10
3. REVIEW OF LITERATURE	12
4. TOOLS AND DATABASES USED	23
5. METHODOLOGY AND APPROACH	35
6. RESULTS	39
7. DISCUSSION	75
8. CONCLUSION	76
9. FUTURE PERSPECTIVES	76
10. REFERENCES	77
11. APPENDIX	79

LIST OF FIGURES

S.No.	Title	Page No.
1.	Embryonic Stem Cell Markers	18
2.	Methodology	35
3.	Graph of TeratoScore	39
4.	CellNet figures showing Heatmap, GRN and NIS Graphs	40-42
5.	PluriTest figures showing Pluriscore, Novelty Score graph	43-44
6.	Interaction network of all marker genes using STRING	56
7.	Final Interaction Network built using STRING	59
8.	Excel sheet of Non-Normalized Microarray Data of GSE72078	60
9.	Excel sheet of Normalized Microarray Data	61
10.	Calculated Threshold	62
11.	Excel Sheet showing matched marker genes present in cell samples	63
12.	Boxplots Showing differences in Normalized and N.N. Data	64
13.	Graphs showing Heatmap of expressed genes in cell sample	67
15.	PluriTest graphs for the validation of our results	73

LIST OF TABLES

S.No.	Title of Table	Page No.
1.	PluriTest Table Showing Pluri raw values	43
2.	Table Showing total 175 marker genes present in stem cells	55
3.	Table acquired from STRING showing various nodes (involved in the interaction network of pluripotency marker genes) and their scores	57-60
4.	Showing Gene list after running java program for finding no. of matched genes in control cell sample.	61
5.	Excel Table showing matched marker genes with their respective expression values and log FC score.	62
6.	Showing Gene list after running java program for finding no. of matched genes in Test cell sample.	63
7.	Table showing range of pluripotent cells, set for the determination of level of pluripotency in any cell type.	69

LIST OF ABBREVIATIONS

ESC	- Embryonic Stem Cells
IPSC	- Induced Pluripotent Stem Cells
Oct-4	- Octamer Binding Transcription Factor 4
KLF4	- Kruppel like Factor 4
SOX 2	- Sex determining region Y, Box-2
STRING	- Search Tool for the Retrieval of Interacting genes
NCBI	- National Centre for Biotechnology Information
GEO	- Gene Expression Omnibus
POU5F1	- Pit Oct Unc domain, class 5 transcription factor 1
Log FC	- Logarithm of Fold Change value
NN	- Non- Normalized data

OBJECTIVE:

To Develop an In silico approach/method which can easily determine Pluripotency of any cell especially iPSC's by using microarray data gene expression profiling in a TEXT file format.

- Comparative analysis of existing tools available for determination of pluripotency.
- Comparative analysis of tools to identify their functional differences.
- Development of a method to find pluripotency of any cell using TEXT file format.

ABSTRACT:

Bioinformatics is proving itself as a boon to modern biological researches; it is serving as a cross-disciplinary field for computational biology solutions to the various problems arising in this domain. Stem cell derived regeneration process is very complex and dynamic concept. Therefore, various statistical and probabilistic aspects are being considered to tackle the situation. Pluripotency, which is itself a big game-changer in the world of Stem cells studies; to get a clear cut idea of how to determine it in any particular cell (either IPSC or ESC) is now a handy deal with the help of various Bioinformatics tools, databases and algorithms. Many researchers have developed several ways to determine Pluripotency but all those tools and methods have their own limitations at their individual level. In this Project we have developed a novel way to determine pluripotency, this method does not rely on any cell type or any special file format acceptance criteria as other available tools do, rather it is providing a simple and reliable way to get access to any cell's pluripotency level. This approach initially gathers the knowledge of working methods by using various other available tools and then developed its own method, which gathers information on genes expression values (present in particular cell) and then several programming languages viz. R, JAVA were used to determine the level of pluripotency. Through this approach we will be able to get a deep knowledge about the level of each gene expression present in any pluripotent cell. In other words, we are able to easily determine which genes are playing more crucial role in making any cell pluripotent. This approach will definitely prove as the most reliable way of determining pluripotency.

Keywords: IPSC, ESC, POU5F1, Nanog, Pluripotency.

INTRODUCTION:

Stem Cells study is a vast topic to discuss about. Several researches are going on to tackle with the major obstacles of this field. Current researches on iPSCs reprogramming methods are on its peak and proving as a future tool for revolutionizing the medical field. The main aspect of stem cell study is pluripotency, as we all know that stem cells could be Unipotent, totipotent, multipotent or pluripotent. Cell differentiation capacity to any of three cell lineages i.e. Ectoderm, endoderm, Mesoderm is dependent upon the potency of that particular cell (Solter, 2006). Embryonic stem cells pluripotency is depend upon the regulation of transcription factors and the epigenetic modifications (Niwa et al., 2000; Mitsui et al., 2003; Chambers et al., 2003; Boyer et al., 2005; Niwa et al., 2005; Boyer et al., 2006). Moreover, further studies also implies that pluripotency of any cell is also a balance nature of some other factors known as Markers. Other than transcription factors these markers also include cell surface markers, pathway related markers and lectins, peptides markers (Wenxiu Zhao et.al. 2012). Today, In vitro approaches are providing a far more reliable path to come over many biological problems; especially bioinformatics tools and softwares are in high demand. Here, in our study we are dealing with such types of tools and databases to cope up with the situation of determining the pluripotency of any cell specially iPSCs.

Many tools have been developed and most of them are available online free of cost. In our study we initially made a comparative analysis of such tools which are working on determination of pluripotency of cells. So, in this race we found 3 main online tools viz. PLURITEST, CellNet, TERATOSCORE, many others are also available there but these three gave the best results and are in high demand. We found that each of them is accepting microarray file data but in different formats e.g. PluriTest accepts only illumina generated .idat* (Raw intensity file) file format, while rest two are using affymetrix generated .cel* file format. But outputs they are providing are far distinct with each other, which we have discussed later in detail. So, by taking this idea we have developed our own approach or method to determine pluripotency by taking TEXT file as input data. Why we use TEXT file? The answer is the vast availability of this file. Text file formats are available for any kind of microarray analysis with their respective default formats. Say for any cell line if .idat format is there but we want to do potency analysis for a particular cell line using PluriTest tool but we don't have the data in .idat format, then we are unable to do so. Also their study is limited to file sizes and quantity (no. of files to be detect) too. But by using our text based approach one can identify pluripotency of any kind of cell line by using its text file format.

For this approach we initially after gone through several literatures, we came to a conclusion that there are total 175 genes known as marker genes present in any pluripotent cell (Wenxiu Zhao et.al. 2012). To validate this conclusion we create an interaction network of our genes using STRING and after filtering our results on the basis of score provided by string, we fetch out top scorer genes.

Our next step was to create a method which can identify pluripotency level of any cell. So, for this we collect microarray data files (Control sample) from NCBI's GEO (Gene Expression Omnibus) and analyze our result with GEO2R (a GEO tool). Then, we download gene expression data file for that particular dataset. After preprocessing step which includes Gene matching (Using JAVA) and data arrangement, we adopt quantile function to calculate Threshold for any cell to be pluripotent. This threshold is based upon the Log FC value (Fold Change value) of the particular gene. After getting threshold we now took another data for test sample and again repeating the same process (as done for control sample), we check for whether the log FC value of test sample passing the threshold or not and we saw that it is passed, as data we have collected for test sample is iPSC data.

Now, our main concern is to check that at what level the cell contains pluripotency? Because the cells which are passing the threshold could either be multipotent or totipotent too. So, to come out of this problem we have developed a JAVA program. This JAVA program is trained by giving it a particular range for particular key regulator gene in either of ESC and iPSC condition. The result of the program is divided into three categories viz. Highly Pluripotent, Partial pluripotent, Low Pluripotent.

Also, to validate our findings and our program, we took data for such cell which is having both .idat* file format as well as .txt file format. So, that we can compare our results with the results of PLURITEST; and we succeed. As the results provided by PLURITEST are much similar to the results provided by our text based approach.

Soon, in near future an online tool could also be developed by which any one can easily access his/ her findings for the determination of their related cell lines on the move of a single click. This approach is a novel work for pluripotency determination as it is providing our own way for creating a method to develop new things using Bioinformatics as a tool.

REVIEW OF LITERATURE:

Stem cells: Cells which have the capability to differentiate into the specialized cell form are termed as Stem cells. These kinds of cells are mostly found in multicellular organisms. In mammals, basically, two different types of stem cells are present: ESCs (Embryonic stem cells), which are extracted from the ICM of blastocysts, and adult SCs, found in various body tissues. During embryonic developmental stage the stem cells are able to differentiate into all the specialized cells lineages i.e. endoderm, mesoderm and ectoderm.

The classical definition states that a cell is said to be stem cell when it possess these two major properties:

- *Self-renewal:* This is the ability of a cell in which it undergoes through several cell division cycle by maintaining its undifferentiating state.
- *Potency:* The potential of any cell to differentiate into any kind of cell lineage is termed as potency. For this the cell must be either totipotent or pluripotent (able to give rise to any mature cell type), although multipotent or unipotent progenitor cells are also referred to as stem cells.

Potency definition

Potency defines the potential of any stem cell for differentiation into any kind of cell lineage.

- Totipotent or omnipotent stem cells can be differentiated into any of embryonic and extraembryonic cell types. Such type of cells is able to develop a complete viable organism.
- Pluripotent SCs are the descendants of omnipotent cells and have the ability to differentiate into almost all the cell lines.
- Multipotent SCs have the potential to differentiate into number of cell types, but only to those which belongs to closely related family.
- Oligopotent stem cells have the ability to differentiate into a few kinds of cell types, such as lymphoid SCs or myeloid stem cells.
- Unipotent stem cells can only differentiate into single cell type, i.e. to their own type but have the potential of self-renewal, which varies them from non SCs.

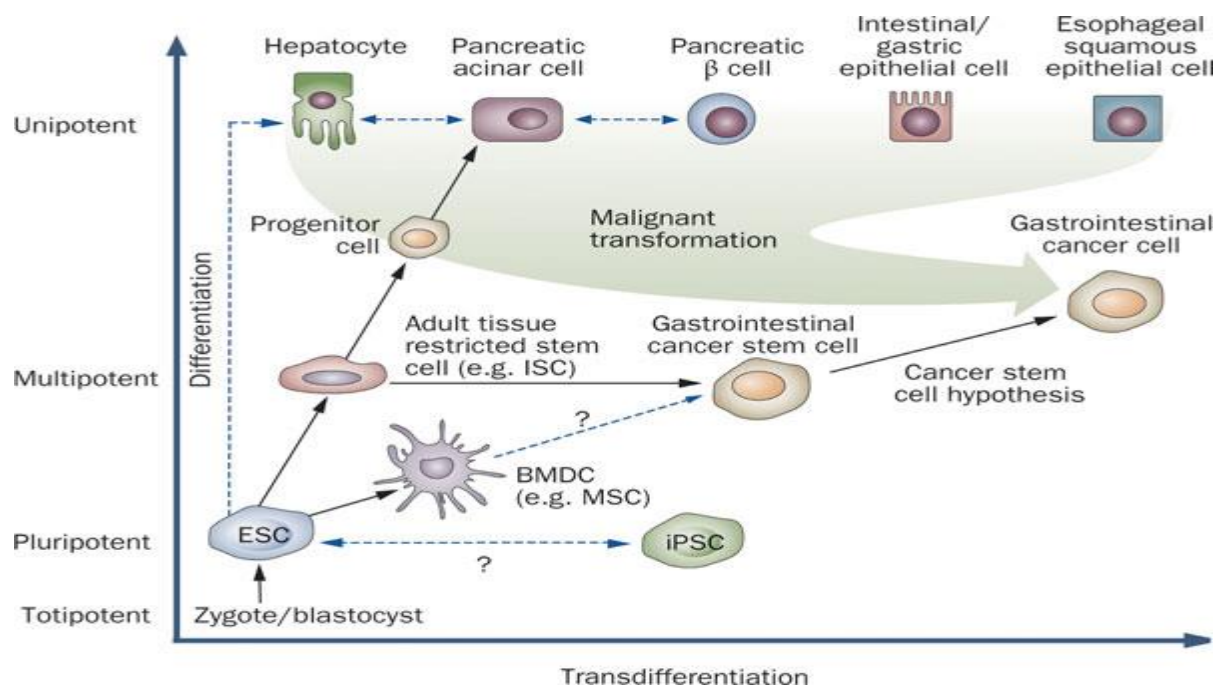
What Are Pluripotent Stem Cells?

There are mainly two types of stem cells present: Adult stem cells and embryonic stem cells. Adult stem cells or somatic cells are found in most of the tissue types. They lose the ability to differentiate into any other cell types, but have the capability to repair tissues which got damaged after cell division and are also helpful in replacing the dying cells. Embryonic stem cells are able to differentiate into any cell type. These cells will later on transform into any of the three germ

layers, this includes endoderm, mesoderm and ectoderm. A stem cell that has the ability to be differentiated into any of these layers is called pluripotent stem cell.

It is now become possible to convert somatic cells into pluripotent stem cells known as reprogramming through genetic engineering. Typically, genes from stem cells known as transcription factors are introduced into somatic cells to reprogram differentiation. The resulting cells are known as induced pluripotent stem cells.

Induced pluripotent stem cells are a type of pluripotent stem cell that can be generated directly from adult cells (Somatic cells). Pluripotent stem cells hold great promise in the field of regenerative medicine. Because they have the capacity to propagate indefinitely, as well as can give rise to most of the cell types in the body (e.g. neural, heart, pancreatic, and liver cells), they represent as only source of cells that could be used in place of those cells which got lost due to damage or disease. The most renowned type of pluripotent stem cell are the ESCs. However, since the generation of embryonic stem cells involves destruction (or at least manipulation) of the pre-implantation stage embryo, there has been much controversy surrounding their use. Further, because embryonic stem cells can only be derived from embryos, it has so far not been feasible to create patient-matched embryonic stem cell lines.

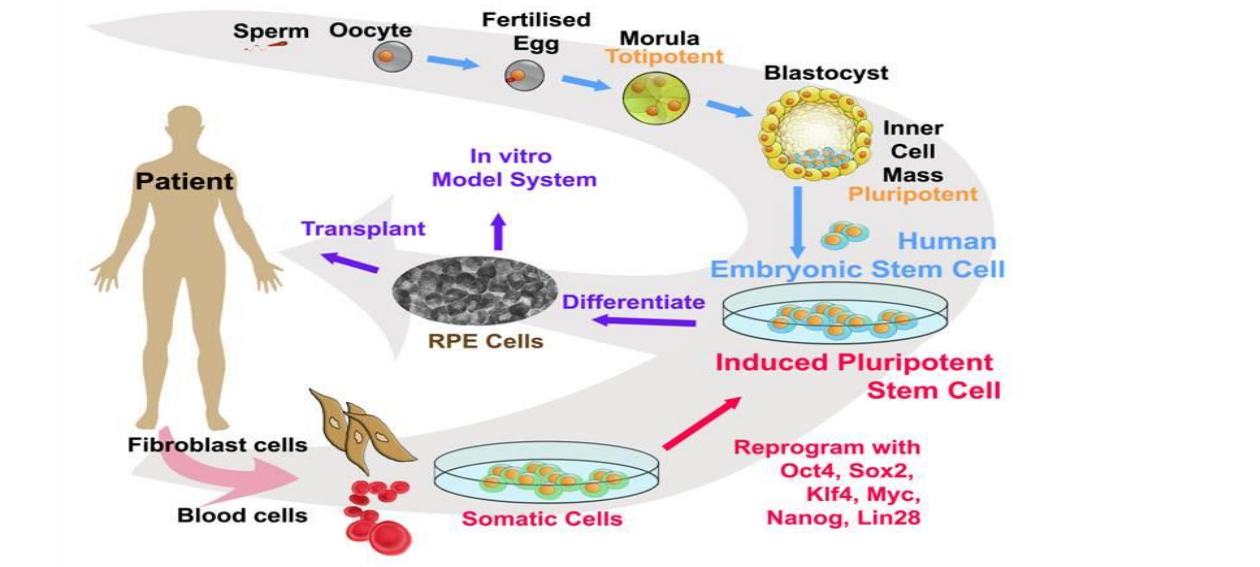


1.1 Previous work:

During the 1950s, Briggs and King established the technique of SCNT, or “cloning,” to probe the developmental potential of nuclei isolated from late-stage embryos and tadpoles by transplanting them into enucleated oocytes. This work, together with some other experiments by Gurdon (Gurdon 1962; Gurdon et al. 1975), showed that differentiated amphibian cells indeed retain the genetic information necessary to support the generation of cloned frogs. The major conclusion from these and subsequent findings was that development imposes reversible epigenetic rather than irreversible genetic changes on the genome during cellular differentiation. The cloning of Dolly the sheep (Wilmut et al. 1997) and other mammals from adult cells, including terminally differentiated cells (Hochedlinger and Jaenisch 2002a; Eggan et al. 2004; Li et al. 2004; Inoue et al. 2005), showed that the genome of even fully specialized cells remains genetically totipotent; i.e., can support the development of an entire organism. However, most of the cloned animal’s possess pungent to severe abnormalities associated with phenotypic and gene expression, suggesting that SCNT results in faulty epigenetic reprogramming (Wakayama and Yanagimachi 1999; Hochedlinger and Jaenisch 2002b; Humpherys et al. 2002; Ogonuki et al. 2002; Tamashiro et al. 2002; Gurdon et al. 2003). To determine transcriptional key regulators that can reprogram adult somatic cells into pluripotent cells, Yamanaka and Takahashi (Tokuzawa et al. 2003) devised a unique scenario for regulatory factors within a pool of 24 pluripotency-associated genes that could activate a drug resistance allele integrated into the ESC-specific *Fbxo15* locus. The combination of 24 factors, when co expressed from retroviral vectors in mouse fibroblasts, indeed activated *Fbxo15* and induced the formation of drug-resistant colonies with characteristic ESC morphology (Takahashi and Yamanaka 2006).

Successive rounds of elimination of individual factors then led to the identification of the minimally required core set of four genes, comprising *Klf4*, *Sox2*, *c-Myc*, and *Oct4*. IPSCs generated by selection for *Fbxo15* activation expressed markers of pluripotent stem cells such as SSEA-1 and Nanog, generated teratomas when injected subcutaneously into immune compromised mice, and contributed to different tissues of developing embryos upon blastocyst injection (Takahashi and Yamanaka 2006).

Generation of iPSC



1.2) Transcription Factors:

There are mainly 4 types of transcription factors which are involved in cellular reprogramming viz. OCT-4, SOX2, NANOG, KLF4, some other are also there which are showing great importance but these are most vital.

1.2.1) Oct-4:

Oct-4 (Octamer-binding transcription factor number 4) also called **POU5F1** (Pit Oct Unc domain class 5 trans. factor 1) is a protein found in human body and is encoded by POU5F1 gene. It is a homeodomain T.F. of the POU family. It is involved in the reprogramming of undifferentiated ESCs, due to which it is also frequently in use as a marker for identification of undifferentiated cells. The octamer (made of eight units) is the nucleotide sequence of DNA "ATTTGCAT" present in the transcription family. The expression of OCT-4 is linked with the expression of undifferentiated phenotype and tumors. Differentiation is promoted by Gene knockdown of Oct-4. Sox2 can form a heterodimer with OCT-4, so that they can bind DNA together. In Mouse embryos it is found that the low expression level of oct-4 leads to failure in formation of inner cell mass, lose in pluripotency and differentiation into trophectoderm.

In 2000, Niwa et al. used conditional repression and expression in murine ESCs to determine the basic requirements of Oct-4 for the maintenance of developmental potency. It is found that optimum level of Oct-4 responsible for three different fates of ESCs. The increase in the expression level of oct-4 up to two fold or less could cause differentiation into either of endoderm or mesoderm. In contrast, Oct-4 repression could leads to loss of pluripotency and trophectoderm dedifferentiation. Thus, an optimum amount of Oct-4 expression is crucial to maintain self-renewal of stem cell as well as the up and down regulation helpful in inducing divergent developments.

The transcription factors such as Oct-4, NANOG and Sox2 are found to be capable of inducing the expression of each other, and are crucial for the maintenance of self-renewing undifferentiated state of the inner cell mass, as well as in ESCs.

1.2.2) SOX-2:

SRY (sex determining region Y)-box 2, a.k.a. **SOX2**, is another T.F. that is found to be essential for the maintenance of self-renewal state or pluripotency state of undifferentiated ESCs. It also has a crucial role in the maintenance of embryonic as well as neuronal stem cells. Sox2 is the member of the Sox family of transcription factors, which is found to be playing crucial role in various stages of mammalian development. This family of protein shares highly conserved domains for DNA binding which is known as HMG (High-mobility group) domains, which contains approx. 80 a.a. It holds a great place in research works, which include IPSCs, which is proving as a promising and emerging sector of regenerative medicines.

Another factor known as NPM1, a different class of transcriptional regulator. It involves in the proliferation of cells, it forms individual complexes with Sox2, Oct4 and Nanog in ESCs. These three important pluripotency factors gave their contribution to pluripotency controlling genes, which are regulated by complex molecular network. Sox2 with oct-4 binds to DNA at non-palindromic sites for the activation of various key transcription pluripotency factors. Surprisingly, regulation of Oct4-Sox2 enhancers can occur without Sox2, likely due to expression of other Sox proteins. However, a group of researchers concluded that the primary role of Sox2 in embryonic stem cells is controlling Oct4 expression, and they both perpetuate their own expression when expressed concurrently.

1.2.3) Klf4:

Kruppel-like factor 4 (KLF4) is a member of the KLF family of transcription factors which helps in the regulation, proliferation, differentiation, apoptosis and differentiated somatic cell reprogramming. Researches also suggest that klf4 in certain types of cancer act as a tumor suppressor, including colorectal cancer.

In ESCs, KLF4 has been shown as a good marker of stem cell like potency indicator. In humans, this protein is encoded by KLF4 gene. It is detected that the klf4 transcription factor which is found to be present at the promoter region of an enzymatic subunit of telomerase (TERT), where it conjugate to form complex with β -catenin subunit.

4) NANOG:

NANOG is another category of T.Fs, which is found to be involved in self-renewal of undifferentiated ESCs. In humans, this protein is encoded by Nanog gene. In Humans NANOG protein which is 305 amino acid long protein having conserved homeodomain motif which is found in close proximity to the nuclear component of the cells. The homeodomain region of Nanog forms DNA binding.

In humans Nanog is consist of three main regions viz. N-terminal region, homeodomain region, and C-terminal regions. Like mus musculus's NANOG, the N-terminal region of human NANOG is also found to be rich in Serine, Threonine and Proline residues, and the C-terminus consist of W repeats. The homeodomain in human NANOG ranges from 95 to 155 a.a. residues long. NANOG act as a key T.F. in ESCs and is thought to be a vital factor in maintaining the pluripotency. NANOG is thought to function in complex with other factors such as Oct-4 and SOX2 to establish ESC like identity. These proteins provide us great area of research study because of their capability to maintain pluripotency.

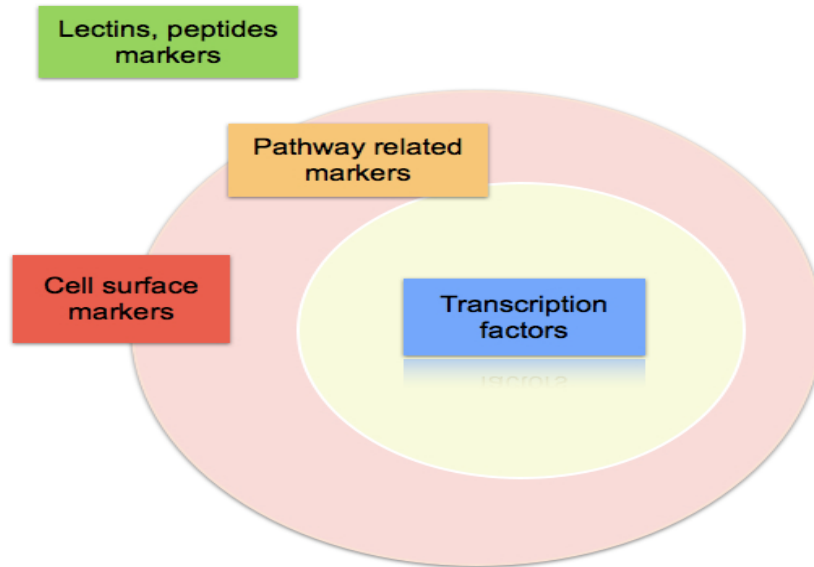
MARKER GENES:

Analysis of arrested embryos demonstrated that embryos express some pluripotency marker genes such as POU5F1, NANOG, Rex1 and many others. Some other specific pluripotency marker genes expression analysis also being captured from human ESC lines:

- TRA-1-60
- TRA-1-81
- SSEA4
- alkaline phosphatase
- TERT
- Rex1

These markers are found to be very beneficial in the identification of pluripotent cells as they allowed for the differentiation in *in vitro* and *in vivo* conditions into the derivatives of all three germ layers viz. ecto., meso., endoderm. POU5F1, TDGF1 (CRIPTO), SALL4, LECT1, and BUB1 etc. related genes, which are also found to be responsible for self-renewal and pluripotency differentiation.

Embryonic stem cell markers



Source : Wenxiu Zhao et.al. – 2012

Identity

Characteristic features of Pluripotent Stem Cells

- On the basis of :
 1. Markers
 2. Telomerase Activity
 3. Differentiation
 4. Methylation
 5. Histone Demethylation

Markers:

1. Cell surface markers:
 - Stage specific Embryonic Antigens.
 - CD Markers.
 - Other Markers.

3. Transcription Factors
4. Signal Pathway related intracellular Markers
5. Enzymatic Markers
6. Markers Overlapping with Tumor Stem Cells

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 are as under:

- **Cellular biological properties:**

- **Morphology:** iPSCs and ESCs are structurally similar to each other. In the way that both have round shape, large nucleolus and scant cytoplasm. Colonies of both of them are also similar. Similar to ESCs, Human iPSCs also forms flat, tightly packed and sharp-edged colonies.
- **Growth properties:** Some cornerstones of ESCs as stem cells must be self-renewed as a part of their standard and these are doubling time and mitotic activity. Some other properties of iPSCs that are prevailing at an equal rate as of ESCs were are that they are mitotically active, self-renewing, proliferating, and dividing at the same rate.
- **Stem Cell Genes:** Some common genes which are found to be equally expressed in both iPSCs and undifferentiated ESCs includes Oct-3/4, Sox2, Nanog, GDF3, REX1, FGF4, ESG1, DPPA2, DPPA4, hTERT, sall4 etc. (Wenxiu Zhao et.al. 2012).
- **Telomerase activity:** Telomeres, which are having ribonucleoprotein heterochromatin like structure which is present at the ends of 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 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.
- **Pluripotency:** iPSCs have capability to differentiate in a fashion as similar to ESCs into fully formed differentiated tissue.
- **Neural differentiation:** iPSCs are also found to be differentiated into neurons, apart from that they are also expresses β -III-tubulin, AADC, DAT, tyrosine hydroxylase, ChAT, LMX1B, and MAP2. The presence of enzymes associated with catecholamine

give indication that iPSCs, which are similar to hESCs, could be differentiated into dopaminergic neurons. The genes associated with Stem cell are usually found to be downregulated after occurrence of differentiation.

- **Teratoma formation:** iPSCs forms 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 proves as a landmark test for the detection of pluripotency.
- **Embryoid body:** Like human ESCs, which forms ball, shaped embryo-like structure also termed as “embryoid bodies”, which consists of all the subjects of mitotically active and differentiating hESCs from all three germ layers. Similarly, iPSCs also form embryoid like bodies and have active differentiated cells.

→ **Epigenetic reprogramming:**

- **Promoter demethylation:**

Methylation referred as the transfer of a methyl group to the base of DNA, especially to the cytosine molecule, which is present in CpG islands (adjacent sequence of cytosine/guanine together in a single stretch). Usually methylation of any gene interrupts with expression by blocking the activity of expression proteins, or by recruiting those enzymes that helps in interfering the expression. Thus, methylation of a gene could be effectively silenced by preventing transcription. Promoters of genes, which are associated with the pluripotency, including Oct-3/4, Rex1, and Nanog, were got demethylated in case of iPSCs.

- **Histone demethylation:**

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 got demethylated, which indicates that the expression of Oct-3/4, Nanog and Sox2.

Traditional Methods of Assessing Pluripotency:

The traditional method includes several techniques as first of them includes the established method for testing cell lines for pluripotency involves injecting stem cells into an animal specimen, usually mice, to observe whether development of Teratomas found or not. Actually Teratomas are tumors that consist of tissues from usually all the three lineages. The mice taken for experiment are usually immunosuppressed, using some genetic mutation process. The

formation of Teratomas in the mice confirms that the considered cell lines are pluripotent type. This method is known as Teratoma technique (Solter et al., 1970; Skreb et al., 1971; Stevens and Little, 1954; Evans and Kaufman, 1981; Stevens, 1958, 1970). Other wet lab based techniques are also there which includes: Chimeras of blastocyst (Tarkowski, 1961; Mintz, 1962), In vitro differentiation assays (Wobus et al., 1984), Transmission of germline (Esteban et al., 2009), Tetraploid complementation assays (Nagy et al., 1993; Kang et al., 2009).

Bioinformatics and computer-based methods for pluripotency determination:

There are many efforts taken by various researchers where they provide reports regarding the development and revolution in the strategy used for the bulk production of human SCs so that they could also be used as regenerative medicine and as well as for the characterization of various pluripotency based assays by providing a number of techniques that promise to improve the unbiased prediction of the uses of both hPSCs and ESCs by using different bioinformatics and gene expression profiling tools. Several online tools, which are doing great in this regard, are discussed below:

PluriTest:

The PluriTest is a diagnostic test based on the DNA microarray. A microarray matches a DNA sequence with its complementary mRNA strand. It essentially identifies genes that are expressed in a cell. The researchers at Scripps have created a microarray database that contains all the genes in embryonic pluripotent stem cells, induced pluripotent stem cells and a few non-pluripotent cells. With this data, they created a model for accessing potency information from normal pluripotent stem cells.

Those who want to determine either their cell lines are pluripotent or not, would have to create a microarray file in *.idat (Raw intensity file) format and upload the data to the website <http://www.pluritest.org>. You need Microsoft Silverlight to access the site, and the file extension for uploaded data is must be .idat. In a relatively short time, the uploaded data is compared to the PluriTest microarray database, and the results of the analysis are displayed. It will include information about pluripotency and any abnormalities in the stem cells (Franz-joseph Muller et. al - 2011)

CellNet:

CellNet is a network biology-based computational platform that more accurately accesses the fidelity of cellular engineering than existing methodologies and generates hypothesis for improving cell derivations. (<http://cellnet.hms.harvard.edu/>)

We can use CellNet online by uploading our data, or we can download and run CellNet locally. We can also use this site to search for predicted transcriptional targets of over 1200 mouse or human transcriptional regulators (Patrick Cahan et. al-2016)

TeratoScore:

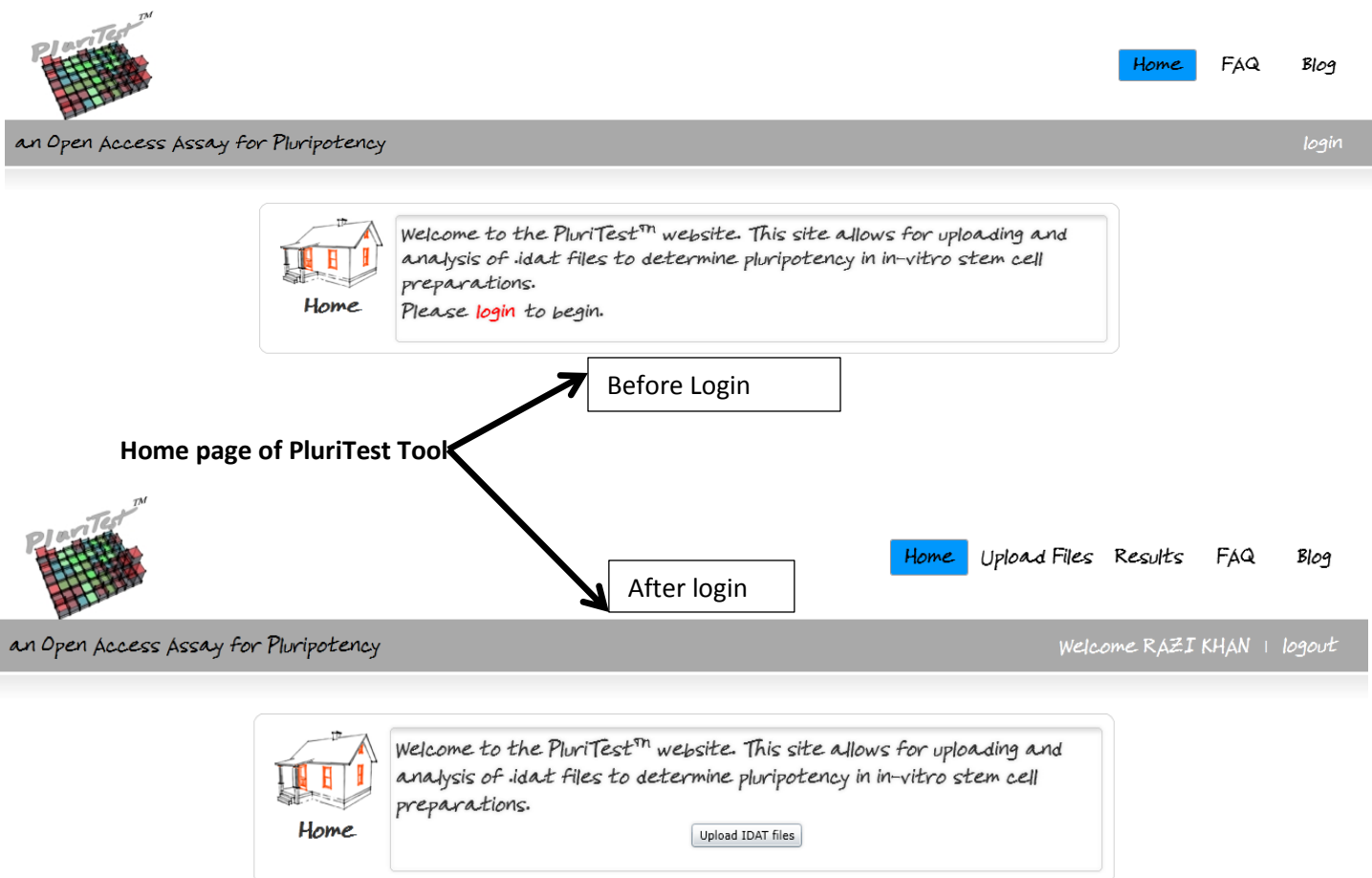
Teratoma formation is the gold standard assay for testing the capacity of human pluripotent stem cells to differentiate into all embryonic germ layers. Using gene expression data from a wide variety of cells, we created a scorecard representing tissues from all germ layers and extraembryonic tissues. TeratoScore, which is an online, open-source platform, distinguishes pluripotent stem cell-derived Teratomas from malignant tumors, by translating cell potency into a quantitative measure (<http://benvenisty.huji.ac.il/teratoscore.php>). The Teratomas used for the algorithm also allowed us to examine gene expression differences between tumors with a diploid karyotype and those initiated by aneuploid cells. Chromosomally aberrant Teratomas show a significantly different gene expression signature from that of Teratomas originating from diploid cells (Yishai Avior et. al - 2015).

TOOLS AND DATABASES USED:


Here, we have a list of various bioinformatics tools and databases, which are being used during this work. This includes several pre available online tools for pluripotency calculation and it also includes the tools, databases and programming languages used for the development of pluripotency finding approach.

1.) PLURITEST:

PluriTest is an open access online tool for the determination of pluripotency of cell line using microarray data obtained from illumina analysis i.e. (.idat* file format only).







Upload Files

IDAT File Uploader

Select chip version: Illumina HT12v4 [Why only these?](#)

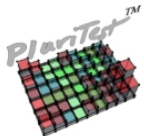
[Select Files](#) [Upload](#) [Clear List](#)


File Name:	GSM2435593_10011068126_D_	2.46 MB Pending
File Name:	GSM2435594_10011068126_E_	2.46 MB Pending
File Name:	GSM2435595_10011068126_L_	2.46 MB Pending

Total Files: 6 0 KB 0%

NOTE: Do not navigate away from this page during the upload process or your analysis will be terminated

Fig.1 .idat* files are uploaded (max. intake is 12 files per analysis)





Results

Completed Analyses

Process Date	# Files	
6/6/2017 11:49:17 PM	6	Details
5/18/2017 5:44:13 PM	6	Details
5/18/2017 1:52:15 AM	6	Details
1/19/2017 10:27:10 PM	5	Details
1/19/2017 9:21:01 PM	6	Details
1/19/2017 9:10:39 PM	12	Details
1/19/2017 9:03:26 PM	4	Details

Times listed are US PST time zone

Analyzed Files

File Name	Pluri Raw	Pluri Logit P	Novelty	Novelty Logit P	PluriTest Result
GSM2435593_10011068126_D_Grn	31.554	1	1.418	0.011	Pass
GSM2435594_10011068126_E_Grn	31.519	1	1.412	0.01	Pass
GSM2435595_10011068126_L_Grn	29.946	1	1.382	0.008	Pass
GSM2435596_10011068132_A_Grn	31.111	1	1.418	0.011	Pass
GSM2435597_10011068132_B_Grn	28.555	1	1.456	0.016	Pass
GSM2435598_10011068132_C_Grn	26.882	1	1.463	0.017	Pass

[View Summary](#)

[View Results](#)

[Results Tables](#)

[View R Log](#)

[View QC Data](#)

[Download Files](#)

[Video](#)

[Output](#)

Fig.2 Result table after analysis

2.) CellNet:

CellNet is a network biology-based computational platform that more accurately assesses the fidelity of cellular engineering than existing methodologies and generates hypotheses for improving cell derivations.

You can use CellNet online by [uploading your data here](#), or you can [download and run CellNet locally](#). You can also use this site to search for predicted transcriptional targets of over 1,200 [mouse](#) or [human](#) transcriptional regulators.

We applied CellNet to expression data from [56 published reports](#). This analysis showed that cells derived via directed differentiation more closely resemble their *in vivo* counterparts than products of direct conversion, as reflected by the establishment of target cell-type GRNs. We have also used the platform to experimentally demonstrate an unanticipated developmental potential of directly converted hepatocytes, and to improve the function of directly converted macrophages. While we have mainly used CellNet in cell engineering, it will also yield insights into the dysregulation of normal transcriptional programs in pathological states, including cancer.

You can now use CellNet to analyze **RNA-Seq data**. You can [download our code to analyze RNA-Seq data or to train a new CellNet platform](#) (e.g. for a different species or to add cell types). See our latest [Nature Protocols manuscript](#) for detailed instructions.

Input → **CellNet** → **Output**

Expression profile → Run locally or on the website → Classification, GRN status, Network influence

CellNet takes as input a gene expression profile and returns:

1. Classification values estimating the likelihood that the profile comes from one of 16 (human) or 20 (mouse) cell- and tissues types. Classification scores are stringent criteria to assess the extent to which an engineered population resembles the training data.
2. Network status, which indicates the extent to which a cell or tissue type GRN is established in the gene expression profile. The GRN status is a sensitive metric of the extent to which specific GRNs are induced or repressed in different conditions.
3. Network influence scores for all transcriptional regulators reflecting the extent to which a transcriptional regulator and its target genes are dysregulated in the query sample, weighted by the importance of the regulator the cell and tissue specific GRN. These scores can be used to prioritize candidate factors to iteratively improve cell engineering.

Fig.3 Showing homepage of CellNet. Click on Run CellNet to continue.

CellNet Run CellNet FAQ About

Analyze your data with CellNet

Please follow the steps listed below to analyze your data with CellNet. You need to upload your raw expression data, upload a sample annotation table, and select several parameters. Typically, the longest part of the run is uploading the data. Once the raw data is uploaded, executing CellNet takes several minutes and you will be notified by email with instructions for downloading the analysis results. The results include normalized data, classification scores, GRN establishment scores, network influence scores, and corresponding figures.

This tutorial contains more detailed instructions on how to use the web app to run CellNet. ([HTML](#), [PDF](#)).

Step 1. Upload your raw expression data

You may upload uncompressed raw expression files (e.g. .CEL files for Affymetrix data). However, to reduce the upload time, we suggest that you compress all of your data files using a compression utility. zip, gzip, or bzip compression formats are accepted. **The maximum upload size of all files together is 128MB.** If your dataset is larger, please [download and run CellNet locally](#). At least 2 .CEL files are required.

No file chosen

Fig. 4 Showing file upload tab with maximum size upto 128MB.

3.) TERATOSCORE:

TeratoScore

translating cell potency into a quantitative measure

Welcome to the TeratoScore web-resource.

In order to inspect teratoma expression pattern, please specify your name and an e-mail address to which you would want to get the analysis to.

Note that TeratoScore currently works with CEL files created on Affymetrix Human Genome U133 Plus 2.0 Array only.

Please enter your name: max 20 chars

Please enter your email:

Please upload .CEL file: No file chosen

[Azrieli Center for Stem Cells and Genetic Research](#)

[The Alexander Silberman Institute of Life Sciences](#)

[The Hebrew University of Jerusalem](#)



Fig.5 Showing Home page of TERATOSCORE having tab for file upload in .CEL format only.

4.) STRING:

(Search Tool for the Retrieval of Interacting Genes/Proteins)

In order to understand cellular processes at system-level as well as at molecular level, Protein-protein interaction networks play an important role. These networks can be utilized for annotating structural, functional and evolutionary properties of proteins by filtering and assessing functional genomics data. Exploring the interaction networks that are already predicted, opens new way for future experimental research and procure cross-species predictions for efficient interaction mapping.

STRING is a biological database of known and predicated protein interactions. The database is freely accessible and updated regularly. The interactions which were derived from four main sources (i.e. Genomic context, High-throughput experiments, (conserved) co-expression and previous knowledge) include physical (direct) and functional (indirect) associations. Therefore, STRING extracts interaction data from these above-mentioned sources and quantitatively distribute for a large number of organisms, and exchanges information among these organisms wherever required.

The latest version 10.5 contains information on about 9.6 millions proteins from more than 2000 organisms. A consortium of academic institutions that include has developed database: Novo Nordisk Foundation Centre for Protein Research, European Molecular Biology Laboratory (EMBL), University of Copenhagen (UCPH), SIB Swiss Institute of Bioinformatics, TU Dresden (abbreviated as TUD) from German, and University of Zurich.

Link for STRING Database: <https://string-db.org/>

The screenshot displays the STRING database homepage. At the top, the STRING logo is on the left, and navigation links for Search, Download, Help, and My Data are on the right. A left sidebar contains a list of search options: Protein by name, Protein by sequence, Multiple proteins (highlighted in blue), Multiple sequences, Organisms, Protein families ("COGs"), Examples, and Random entry. The main content area is titled "SEARCH" and "Multiple Proteins by Names / Identifiers". It features a "List Of Names:" input field with the text "oct4", "sox2", "Nanog", and "klf4". Below this is a checkbox for "... or, upload a file:" with a "Browse ..." button. An "Organism:" dropdown menu is set to "Homo sapiens". A large blue "SEARCH" button is at the bottom.

Fig.6 Showing STRING Homepage with various attributes and Search tab options.

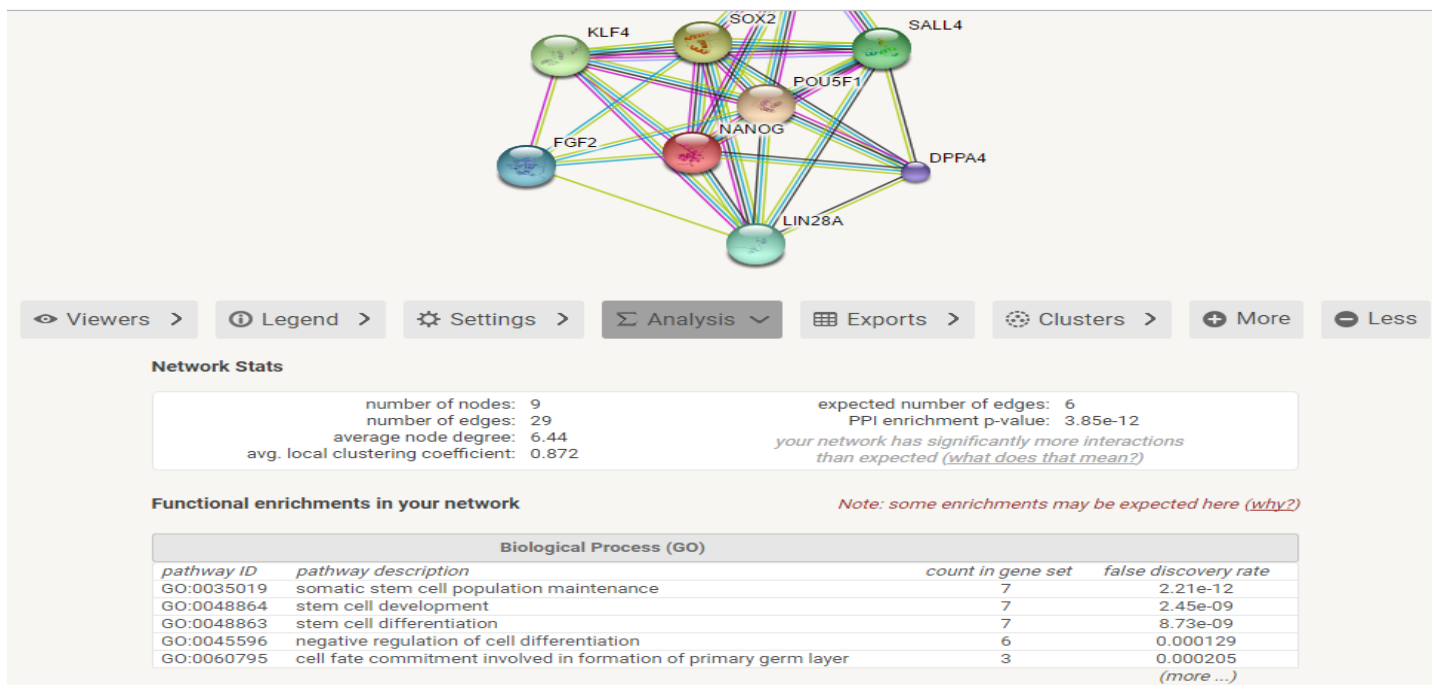


Fig.7 Showing Analysis report of selected/ searched genes or proteins including their interaction network.

5.) GEO (Gene Expression Omnibus):

The Gene Expression Omnibus (GEO, created in 2000) at National Center for Biotechnology Information (NCBI) is a freely available repository that contains microarray data, next-generation sequencing (NGS), and other forms of high-throughput functional genomics data sets. In this database, various tools are provided in order to help users for query and download experiments and curated gene expression profiles. The 90% data in GEO database based on gene expression profiles that look into a broad spectrum of biological themes like disease, development, evolution, immunity, ecology, toxicology, and metabolism. Remaining 10% data in GEO about other categories of functional genomics and epigenomic studies covering those that examine genome methylation, structure of chromatin, copy number variations in genome and genome-protein interactions. Data in GEO is a collection of original research submitted by scientific community with a journal agreement that require data to be made available in a public repository, and the purpose of this to facilitate evaluation of results, reanalysis and full access to all parts of the study.

At present, GEO stores more than a billion individual gene expression measurements that are derived from over hundred organisms, submitted by more than 1500 laboratories all over the globe addressing a wide range of biological phenomena. Several user-friendly web-based interfaces and apps have been developed that enable effective exploration, query, and visualization of these data, at the level of individual genes or entire studies. Link for GEO database: <http://www.ncbi.nlm.nih.gov/geo>.



Fig.8 Showing Home page of NCBI with search tab.

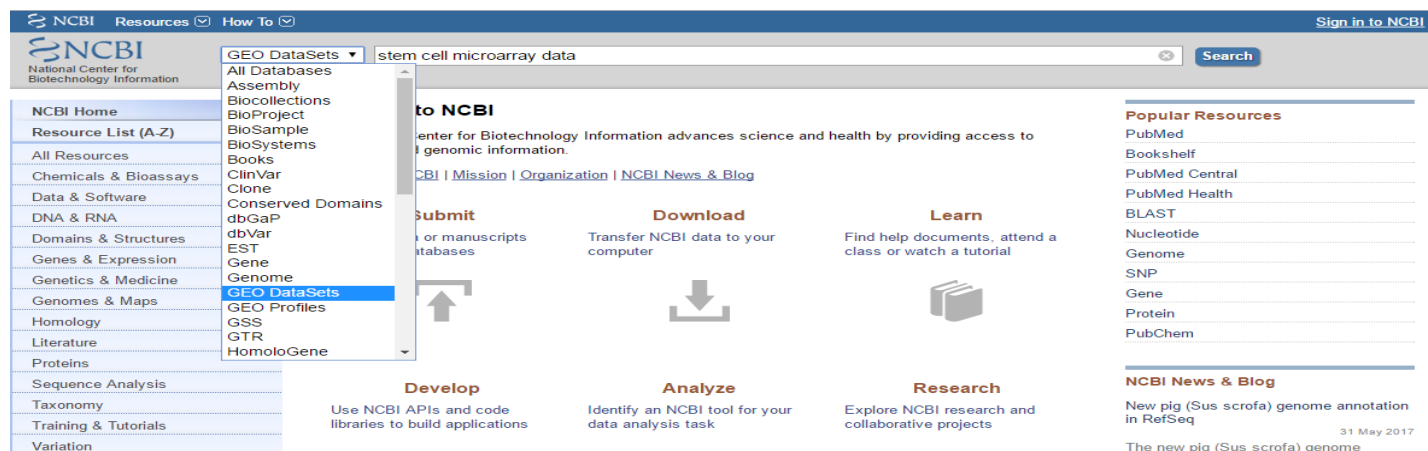


Fig. 9 Showing selection of GEO datasets to get microarray analysis query.

6.) GEO2R (GEO-tool):

GEO2R is a web-based tool of Gene Expression Omnibus database that permits users to compare two or more groups of samples in a GEO series in order to find those genes that are differentially expressed across experimental conditions. Results are obtained in table form, in which genes are arranged according to their significance. The comparison done by GEO2R tool with the help of two R packages of Bioconductor project: GEOquery and limma.

Bioconductor is an open source software project consisting of various packages based on the R programming language that provides basis for the analysis of microarray based high-throughput genomic data. The GEOquery R package analyzes GEO data into R data structures so that it can be used by other R packages. The *limma* (Linear Models for Microarray Analysis) R package has become one of the most widely used statistical tests for identifying differentially expressed genes.

☐ [Gene expression profiles in human ES cells and human iPS cells](#)

3. (Submitter supplied) Genome editing research of human ES/iPS cells has been accelerated by clustered regularly interspaced short palindromic repeats/CRISPR-associated 9 (CRISPR/Cas9) and transcription activator-like effector nucleases (TALEN) technologies. However, the efficiency of biallelic genetic engineering in transcriptionally inactive genes is still low, unlike that in transcriptionally active genes. To enhance the biallelic homologous recombination efficiency in human ES/iPS cells, we performed screenings of accessorial genes and compounds. [more...](#)

Organism: Homo sapiens
Type: Expression profiling by array
Platform: GPL16699 6 Samples
Download data: TXT
Series: Accession: GSE69653 ID: 200069653

[Analyze with GEO2R](#)

☐ [Effect of FGFC on human embryonic stem cells](#)

4. (Submitter supplied) Fibroblast growth factors (FGFs) are essential for maintaining self-renewal in human embryonic stem cells and induced pluripotent stem cells. Recombinant basic FGF (bFGF or FGF2) is conventionally used to culture pluripotent stem cells; however, because of bFGF instability, repeated addition

Fig. 10 For GEO2R Analysis click on the link of *Analyze with GEO2R* (present below the desired dataset).

NCBI

Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO | Login

NCBI » GEO » GEO2R » GSE69653

Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance. [Full instructions](#) [YouTube](#)

GEO accession: [Gene expression profiles in human ES cells and human iPS cells](#)

Selected 0 out of 6 samples

Group	Accession	Source name	Cell line
-	GSM1705351	Tic,undifferentiated	Tic
-	GSM1705352	YOW-IPS,undifferentiated	YOW-IPS
-	GSM1705353	Dotcom,undifferentiated	Dotcom
-	GSM1705354	KhES1,undifferentiated	KhES1
-	GSM1705355	H9,undifferentiated	H9
-	GSM1705356	H1,undifferentiated	H1

GEO2R | Value distribution | Options | Profile graph | R script

Fig.11 For GEO2R analysis define groups of the available cell lines at least 2 groups must be defined.

GEO accession

GSE69653

Set

Gene expression profiles in human ES cells and human iPS cells

▼ Samples

▼ Define groups

Selected 6 out of 6 samples

Enter a group name:

List

Cancel selection

ipsc (3 samples)

esc (3 samples)

Group	Accession	Source name	Cell line
ipsc	GSM1705351	Tic,undifferentiated	Tic
ipsc	GSM1705352	W-IPS	YOW-IPS
ipsc	GSM1705353	Dotcom,undifferentiated	Dotcom
esc	GSM1705354	KhES1	KhES1,undifferentiated
esc	GSM1705355	H9	H9,undifferentiated
esc	GSM1705356	H1	H1,undifferentiated

Fig.12 Cell lines were distributed among groups

esc	GSM1705354	KhES1	KhES1,undifferentiated	KhES1
esc	GSM1705355	H9	H9,undifferentiated	H9
esc	GSM1705356	H1	H1,undifferentiated	H1

GEO2R

Value distribution

Options

Profile graph

R script

▼ Quick start

- Specify a GEO Series accession and a Platform if prompted.
- Click 'Define groups' and enter names for the groups of Samples you plan to compare, e.g., test and control.
- Assign Samples to each group. Highlight Sample rows then click the group name to assign those Samples to the group. Use the Sample metadata (title, source and characteristics) columns to help determine which Samples belong to which group.
- Click 'Top 250' to perform the calculation with default settings.
- Results are presented as a table of genes ordered by significance. The top 250 genes are presented and may be viewed as profile graphs. Alternatively, the complete results table may be saved.
- You may change settings in Options tab.

How to use

Top 250

Save all results

Fig. 13 Save all the results using the below tab and copy the results to spreadsheet.

► Samples

► Define groups

Selected 6 out of 6 samples

GEO2R

Value distribution

Options

Profile graph

R script

► Quick start

Log-transformation has been applied to the data. You can change this in the Options tab.

Recalculate if you changed any options. Save all results Select columns

ID	logFC	GENE_SYMBOL	GENE_NAME
► 25869	-11.355	RPS4Y1	ribosomal protein S4, Y-linked 1
► 41251	-11.911	RPS4Y2	ribosomal protein S4, Y-linked 2
► 44871	2.6	LOC100506389	uncharacterized LOC100506389
► 10436	-1.98	XLOC_003650	
► 10343	2.973	ERAP1	endoplasmic reticulum aminopeptidase 1
► 62583	-1.885	XLOC_004653	
► 51244	-8.192	LOC441666	zinc finger protein 91 pseudogene

Fig.14 Showing Results after analysis.

☐ [Gene expression profiles in human ES cells and human iPS cells](#)

3. (Submitter supplied) Genome editing research of human ES/iPS cells has been accelerated by clustered regularly interspaced short palindromic repeats/CRISPR-associated 9 (CRISPR/Cas9) and transcription activator-like effector nucleases (TALEN) technologies. However, the efficiency of biallelic genetic engineering in transcriptionally inactive genes is still low, unlike that in transcriptionally active genes. To enhance the biallelic homologous recombination efficiency in human ES/iPS cells, we performed screenings of accessorial genes and compounds. [more...](#)
Organism: Homo sapiens
Type: Expression profiling by array
Platform: GPL13399, 6 Samples
Download data: TXT
Series Accession: GSE69653 ID: 200069653
[Analyze with GEO2R](#)

☐ [Effect of FGFC on human embryonic stem cells](#)

4. (Submitter supplied) Fibroblast growth factors (FGFs) are essential for maintaining self-renewal in human embryonic stem cells and induced pluripotent stem cells. Recombinant basic FGF (bFGF or FGF2) is conventionally used to culture pluripotent stem cells; however, because of bFGF instability, repeated addition of fresh bFGF into the culture medium is required in order to maintain its concentration. In this study, we

Ass with

Des Var

App The

Import: GEO Hc

GEO Dc

About G

Fig.15 for downloading Expression values for each gene present in cell lines dataset, download Series matrix file by clicking on download data tab as shown in figure.

NCBI

GEO

Gene Expression Omnibus

NCBI » GEO » Download data

GEO Publications

FAQ

MIAME

Email GEO

Login

Download data for GSE69653

Information about GEO data organization is provided in the [GEO overview](#). GEO FTP site structure and available formats are described in [README.txt](#). Additional download options are described at [Download GEO data](#).

Series SOFT file:

SOFT family files are text files that incorporate complete data and metadata for all Platform, Sample and Series records in the family.
[GSE69653_family.soft.gz](#) 12.7 Mb

Series MINIML file:

MINIML family files are XML files that incorporate complete data and metadata for all Platform, Sample and Series records in the family.
[GSE69653_family.xml.tgz](#) 12.7 Mb

Series Matrix file:

Series_matrix files are text files that include a tab-delimited value-matrix table generated from the 'VALUE' column of each Sample, headed by Sample and Series metadata. These files are suitable for loading into spreadsheet applications such as Excel.
CAPTION: data are extracted directly from the original records with no consideration as to whether the values are directly comparable.
[GSE69653_series_matrix.txt.gz](#) 2.1 Mb

Series supplementary data archive (contains TXT files):

[GSE69653_RAW.tar](#) 18.5 Mb

Fig.16 Showing link for downloading series matrix TEXT file.

7.) R programming language:

R is a programming language, developed at Bell Laboratories, USA by John Chambers and colleagues that provides an environment or platform for statistical analysis and graphics. R comes under GNU project that is similar to S language. Maximum codes written for S, easily run in R. R covers a broad area in statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering) and graphical field and highly flexible.

The main advantage of using R is the ease with which well-designed publication-quality graphs can be obtained, that includes mathematical symbols and different types of formulae where required and user has full control.

R is available as free software under the terms of the Free Software Foundation's GNU General Public License in source code form and it is designed so well that it can be easily compiled and run on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

RStudio is a freely available IDE (integrated development environment) for R language and developed by RStudio, Inc. and Hadley Wickham, at present who are Chief Scientist at RStudio. It is available in two editions: one for desktop, where the program runs locally as a regular desktop application and second is RStudio Server that allows accessing RStudio using a web browser but it is compatible for remote Linux server. RStudio is written in C++ and uses the Qt framework for its graphical user interface. The beta-version of RStudio was officially out in February 2011 and version 1.0 released on 1 Nov 2016.

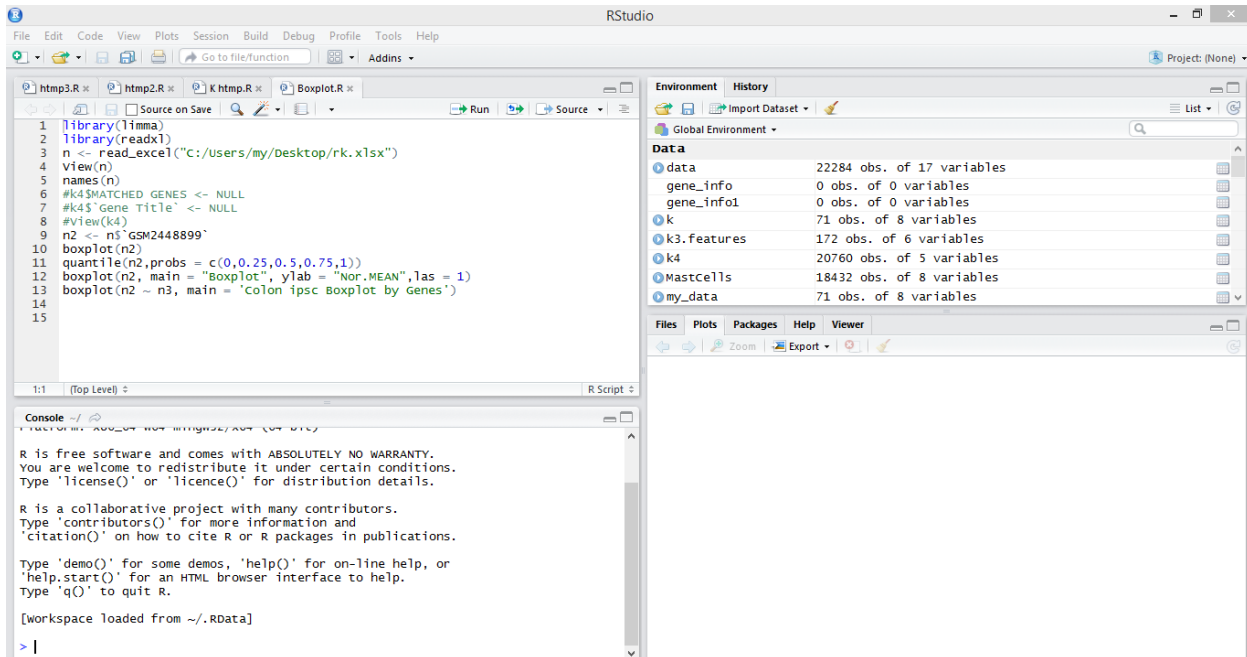


Fig.17 Figure showing R Studio IDE environment for R Language consisting of 4 main divisions: 1. Source Panel 2. Console Panel 3. History panel 4. Export or info panel.

8.) NETBEANS:

NetBeans is a software development platform written in Java. The NetBeans Platform allows applications to be developed from a set of modular software components called *modules*. Applications based on the NetBeans Platform, including the NetBeans integrated development environment (IDE).

NetBeans IDE lets us quickly and easily develop Java desktop, mobile, and web applications, as well as HTML5 applications with HTML, JavaScript and CSS. The IDE also provides a great set of tools for PHP and C/C++ developers. It is free and open source and has a large community of users and developers around the world.

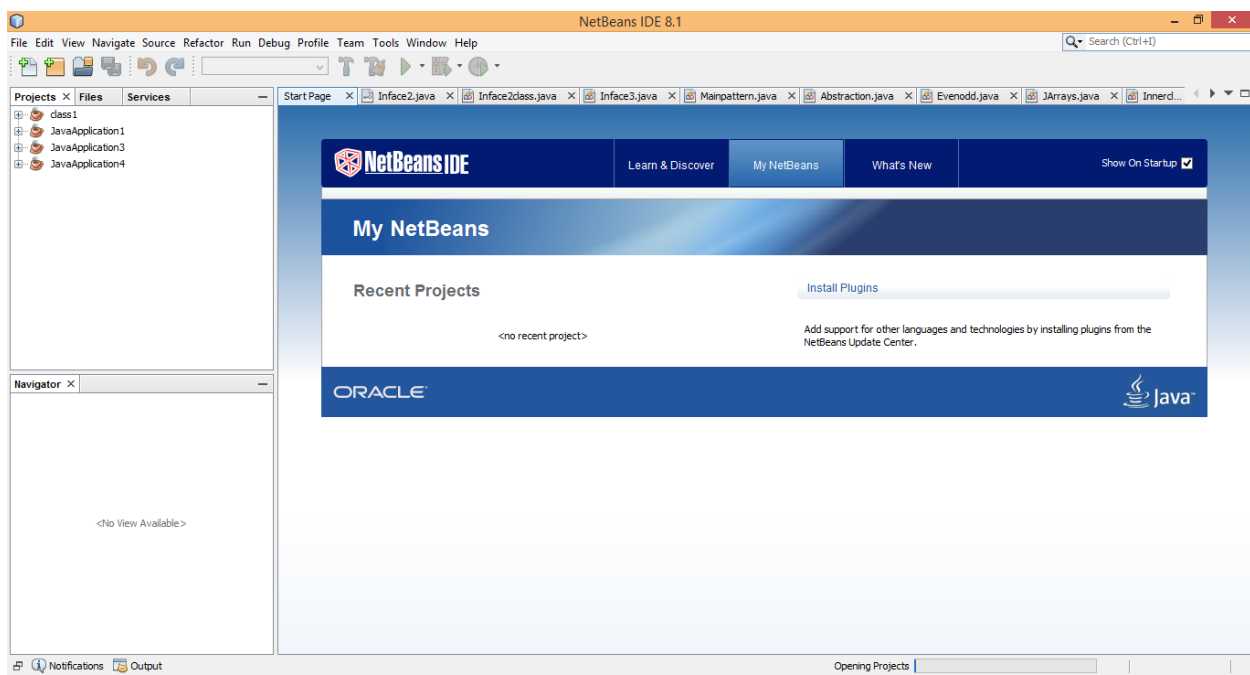


Fig.18 Showing Homepage of NETBEANS IDE for the development of JAVA programs.

METHODOLOGY:

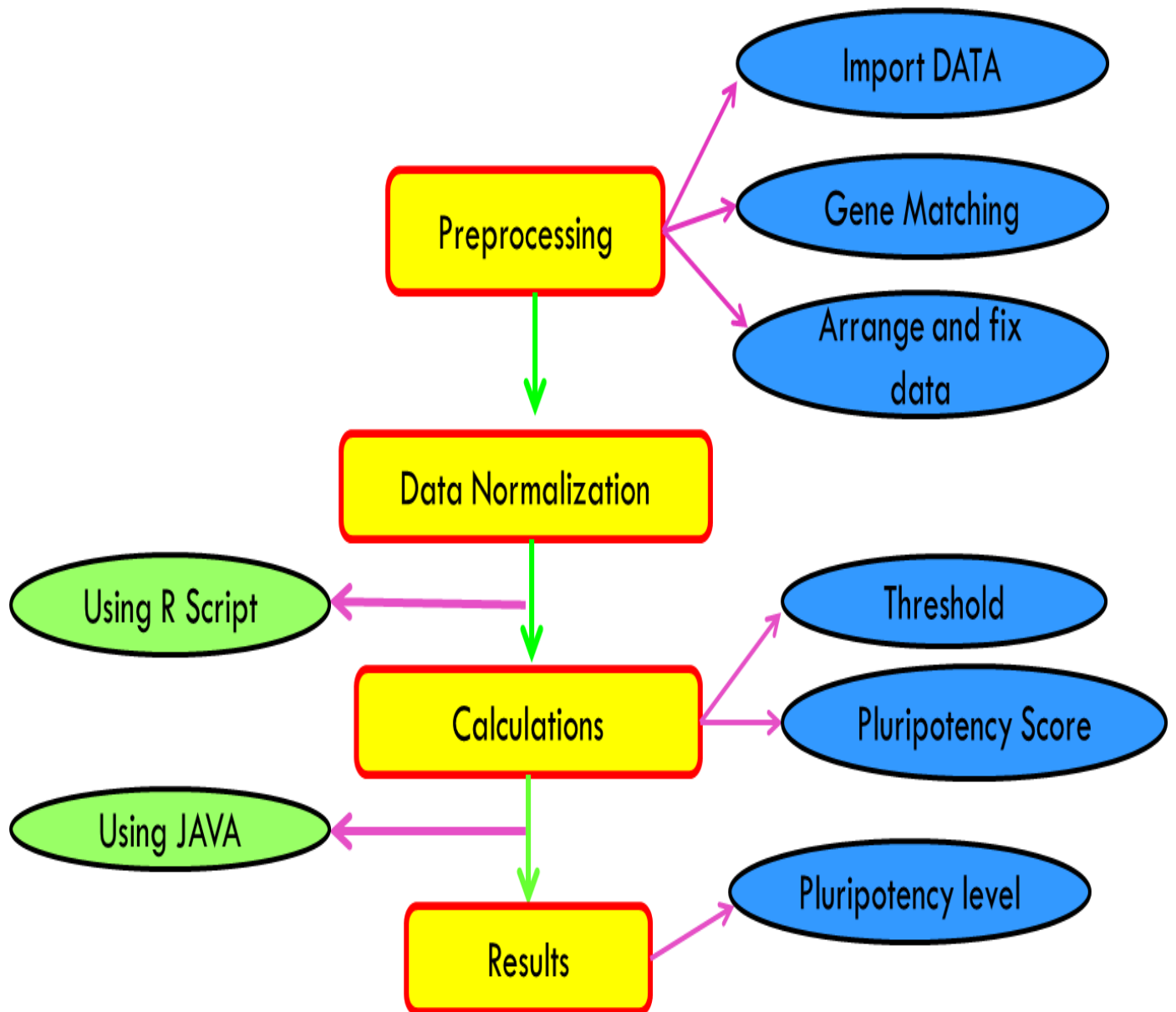


Fig.1 Defining various stages used in this method for finding pluripotency of any cell. The detailed approach is as under.

APPROACH:

1) Comparative analysis of existing tools available for determination of pluripotency.

Step 1: Searching of the tools responsible for the detection of pluripotency in Stem cells.

Step 2: Checkout the working and analysis plan of Each and every tool we have found.

Step 3: Came out with various results provided by the tools.

Step 4: Analyze each and every Result by comparing them at various key points.



2) Development of a method to find pluripotency of any cell using TEXT file format.


Step 1: Gone through several research papers to identify the genes responsible for the pluripotency of the cell (Total 175 Marker genes were found).




Step 2: To validate our findings STRING tool was used to create interaction network between all 175 genes, from here we got probability scores which were satisfactory and ensure us their presence.




Step 3: Filtration was done to get most prominent sets of genes or the key regulators, which later on helpful in determining level of pluripotency.




Step 4: After validation our next step is to set up Threshold value for the cells to pass the pluripotency limit. For this we have taken microarray data of human's whole genome IPSC's (GSE72078) from GEO datasets of NCBI.




Step 5: Next step is preprocessing of the data by arranging them on the basis of matched marker genes with their respective ref IDs and expression values according to the given cell lines(to get matched marker genes JAVA program was developed).




Step 6: After comparison sort the data and normalize it by using any of global Median normalization or Simple Normalization to convert our raw data to appropriate linearly arranged data.




Step 7: R script was used to set threshold by using Quantile function.




Step 8: Now to test our set threshold, a predetermined IPSC test sample (GSE93228) was downloaded and after preprocessing steps, its pluripotency score was calculated to check whether it is passing the threshold or not (Sample pass).



Step 9: Different graphs were plotted viz. boxplots (to compare normalized and non-normalized data), Clustering graphs, Heatmaps etc. using R script.



Step 10: Now, for calculating the level of pluripotency we set a range by collecting expression data of various ESCs and IPSCs and further dividing them into different parameters to get promising results.



Step 11: Range was set for key regulators for ESC as well as for iPSCs independently.



Step 12: Three levels were generated and the ranges were distributed into each level in particular order.



Step 13: Again a JAVA program was developed to determine the level of pluripotency by considering all ranges at different levels.



Step 14: Control and Test sample were analyzed to determine their pluripotency level.



Step 15: To validate our method several other Datasets were checked and the results were compared with the results of PluriTest (Both results are validating each other).



RESULTS AND DISCUSSION:

1) Comparative analysis of existing tools available for determination of pluripotency:

- 1) Results from different available online tools are shown which determines us the feature based aspects of these tools. Also in these results we can see the variations of different tools in accepting and importing data, their individual way of calculations and presenting results for pluripotency. These tools helped us in adapting an idea relies upon the access of pluripotency of any cell line present in Gene expression Omnibus (GEO) with Text file format. So, the results of online tools are as under:

1. TERATOSCORE :

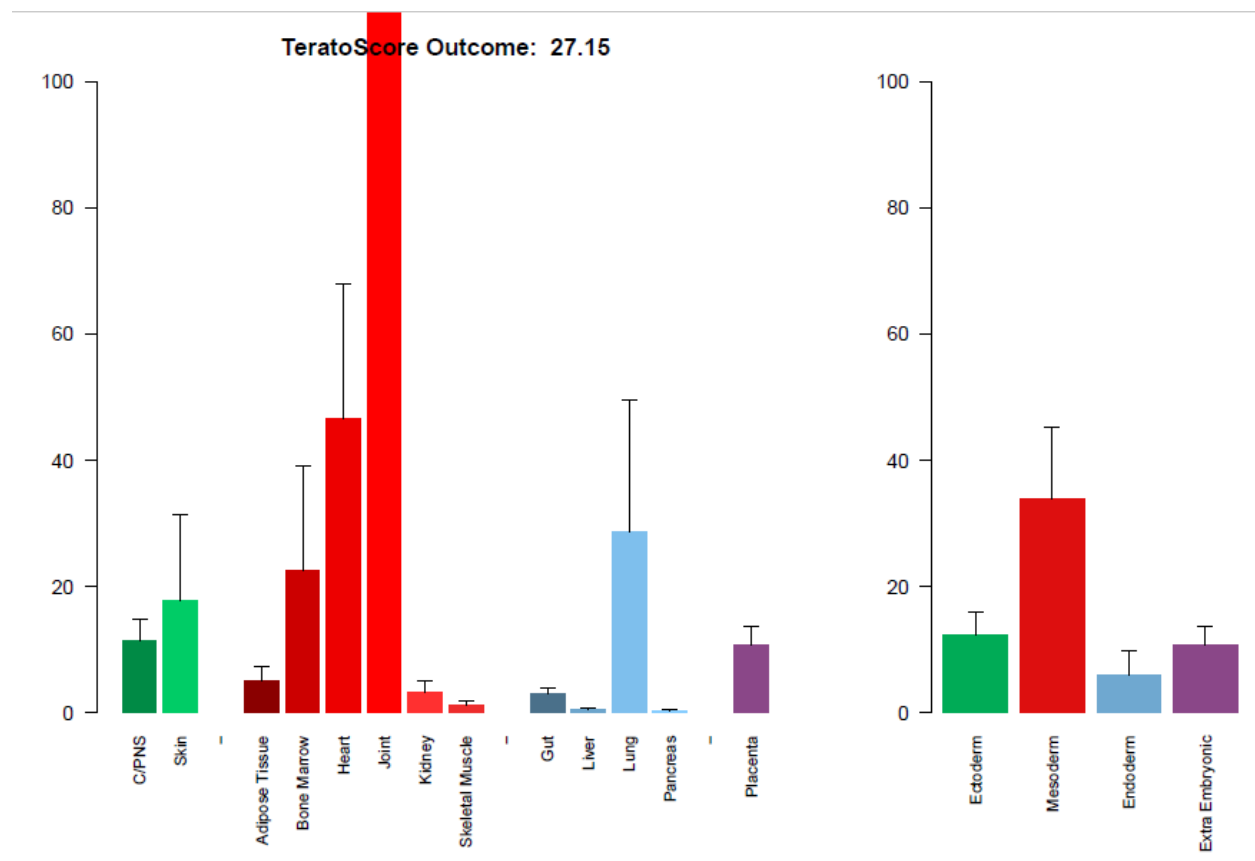


Fig.2 Graph Showing TERATOSCORE outcome of 27.15 for fibroblast cells and comparison with other lineages.

2. CELLNET :

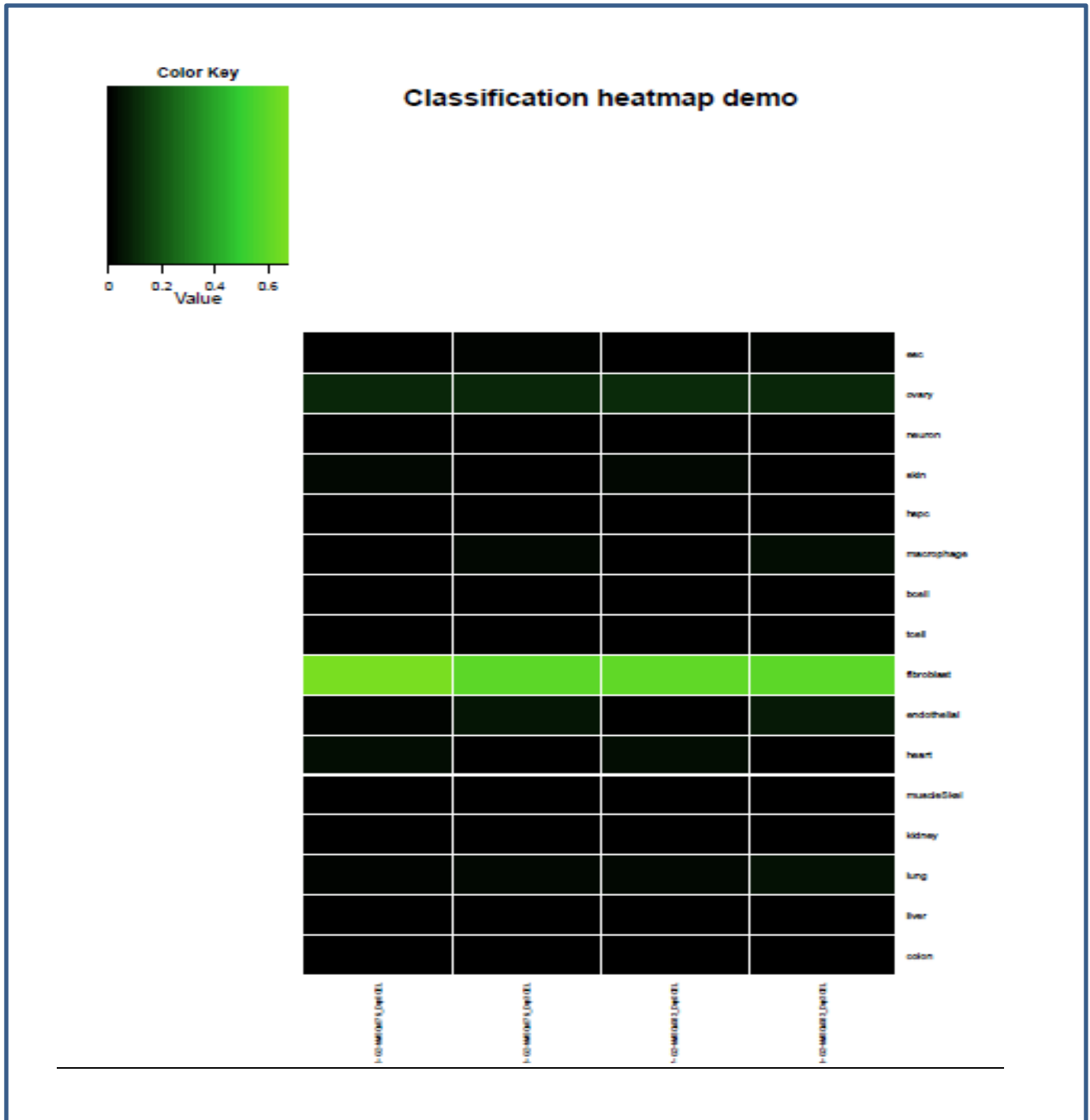


Fig.3 Showing CellNet Heatmap results for fibroblast cells, where genes are highly expressed as compared to other cell lines.

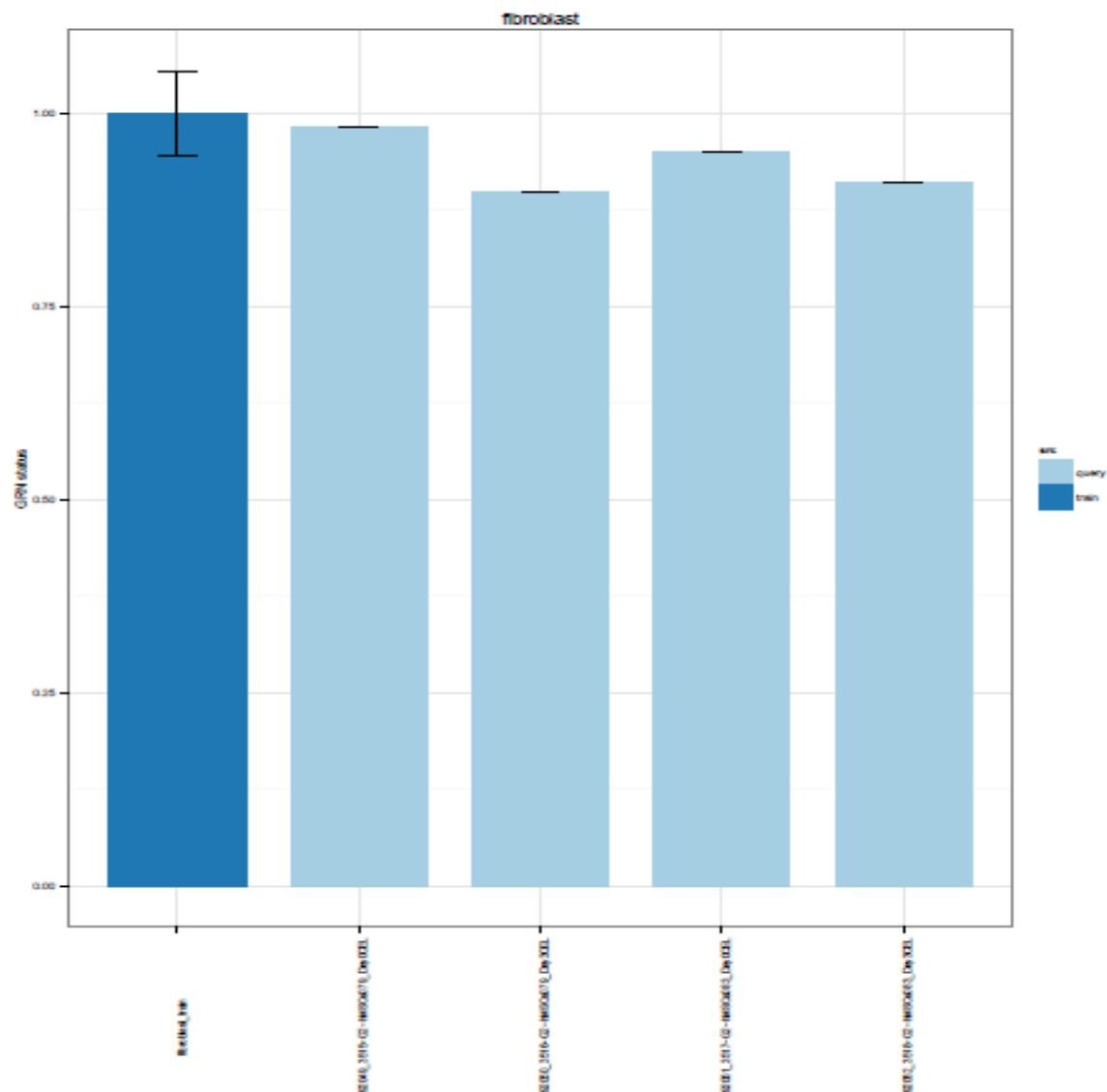


Fig.4 Showing is a data table in which each row represents one GRN associated with a particular cell or tissue type, and each column represents one of your samples. The values represent the extent to which the GRN is established to a level equivalent to that seen in the associated cell or tissue type. The file is named grnScores.csv in your results. We represent this as a bar plot, where replicates are combined into one bar (colored light blue), and the training data of the starting and target cell types are also shown as points of reference (dark blue).

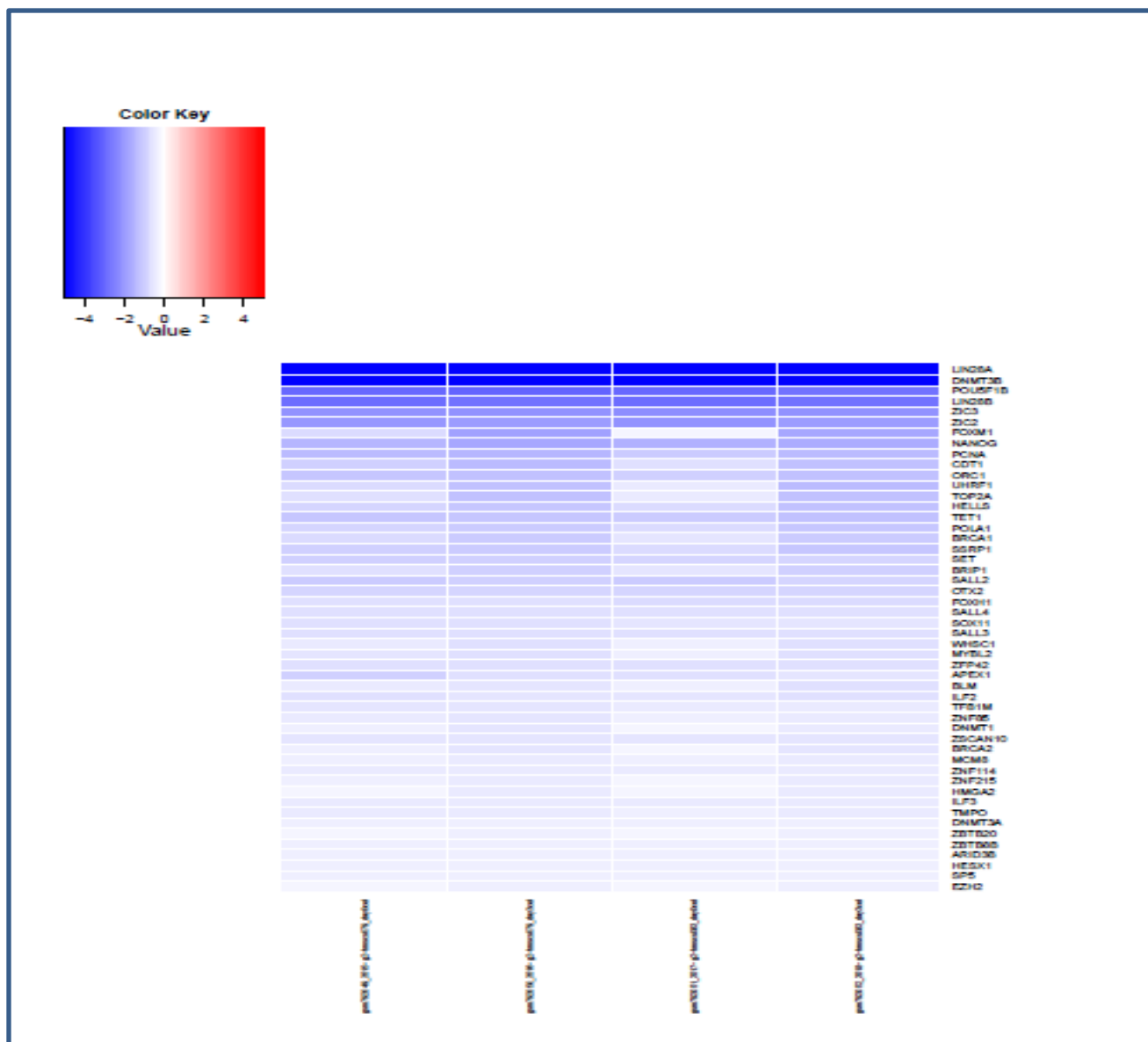


Fig.5 Showing the Network influence score which is computed for each transcriptional regulator in the target cell type GRN according to the extent that it is either too highly or too lowly expressed in your terminal sample. The NIS also integrates other information into scoring transcriptional regulators, including the extent to which predicted target genes are dysregulated, the number of transcriptional targets, and the expression level of the regulator in the target cell type. Positive values indicate that the regulator is too highly expressed, and negative values indicate that the regulator is too lowly expressed.

3. PLURITEST :

Table with results

	pluri-raw	pluri logit-p	novelty	novelty logit-p	RMSD
GSM1695209_10002121031_A_Grn	37.34	1.00	1.35	0.01	0.42
GSM1695210_10002121031_B_Grn	37.64	1.00	1.21	0.00	0.41
GSM1695211_10002121031_C_Grn	25.81	1.00	1.67	0.12	0.52

Fig.6 Table Showing RAW data and Calculated Pluripotency and Novelty score for 3 different *idat format Cell lines of Fibroblast microarray data using Variance Stabilizing Transformation and Loess Normalization by Using PLURITEST Tool.

Model-Based Multi-Class Pluripotency Score

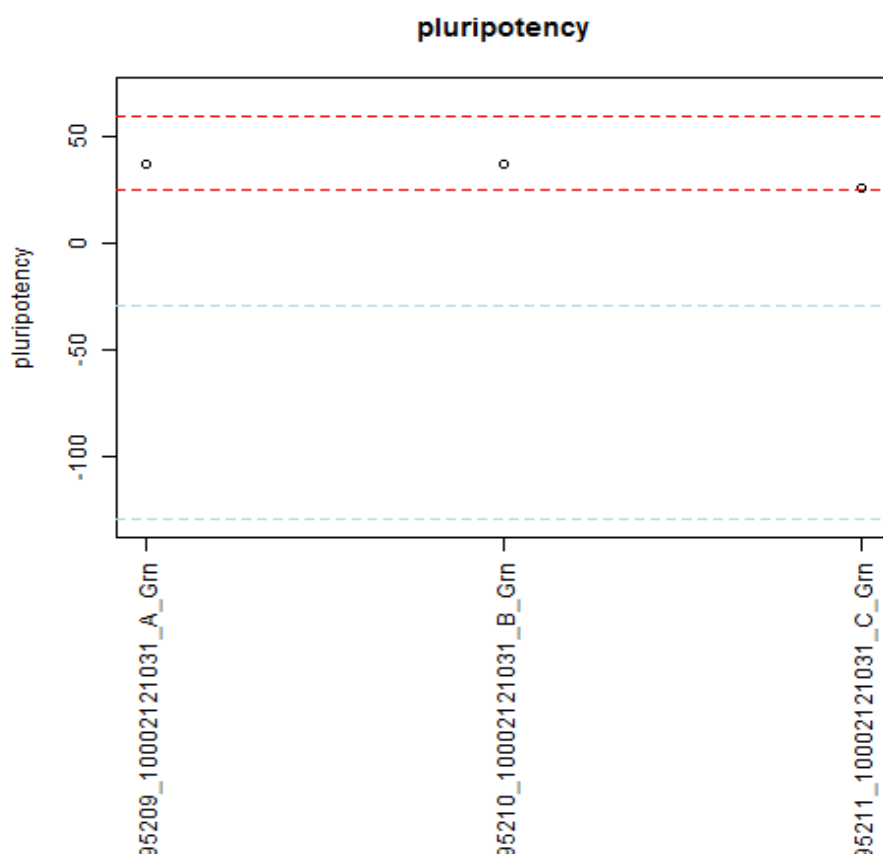


Fig.7 PLURITEST Graph showing pluripotency level of 3 cell lines, which depicts that all three cells are coming under pluripotency threshold set value. Hence, are said to be pluripotent.

Novelty

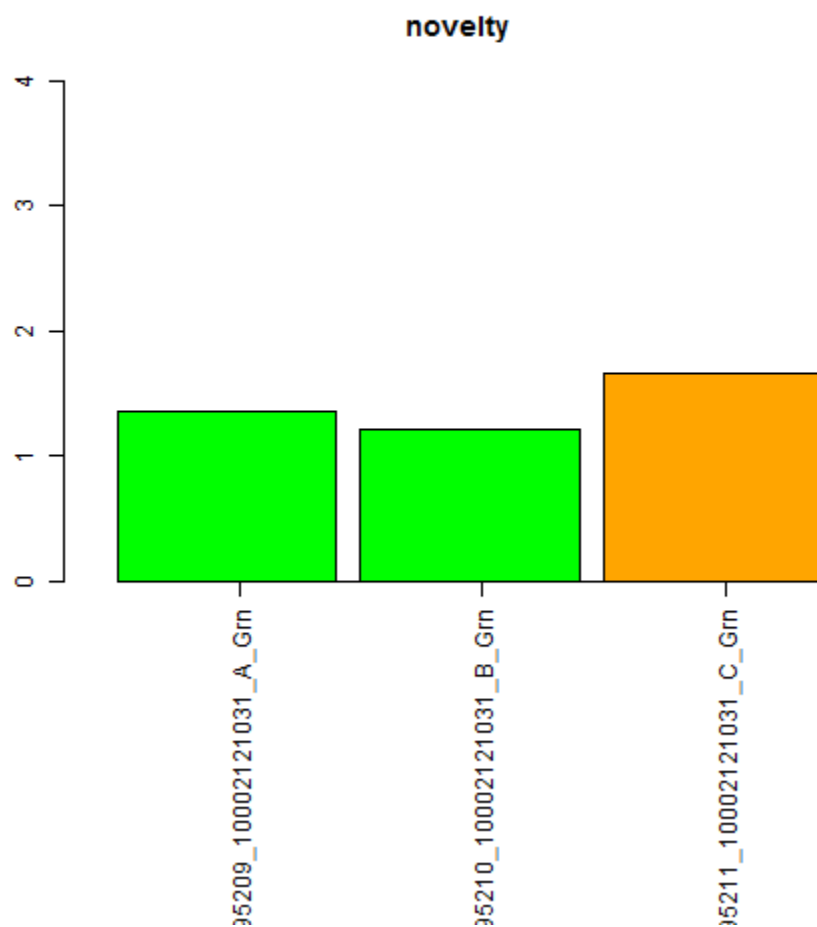


Fig.8 PLURITEST result showing Novelty score, which represents the number of novel genes present in that particular cell line which making it less pluripotent as compared to the training data of PLURITEST. The green color represents the pluripotent state, whereas orange color represents the non-pluripotent or novel state of cell line.

2. Comparative Analysis of tools to identify their functional differences:

Comparative Analysis:

1. On the basis of File type or Microarray Analysis :

This includes the comparison of tools based on platforms of which they are accepting the files. E.g., PluriTest takes only illumina generated .idat* files, whereas CellNet accepts only Affymetrix based .CEL* file formats of following platforms : Affymetrix Mouse Genome 430 2.0 Array (GPL1261), Affymetrix Mouse Gene 1.0 ST Array, Illumina MouseRef-8 v2.0 expression bead chip, Affymetrix Human Genome U133 Plus 2.0 Array, Affymetrix Human Gene 1.0 ST Array. TeratoScore also accepts the file formats of Affymetrix but it uses only Human Genome U133 plus 2.0 platform.

2. On the basis of Algorithms:

All these tools have different algorithm for the analysis of results as in the case of CellNet, it uses Context Likelihood of Relatedness algo. First, they had generated the complete training data that would serve to estimate the expression distributions of genes from each platform. Second, they reconstruct gene regulatory networks. Finally, they use the GRNs to train cell and tissue type classifiers.

In PluriTest, our uploaded file is preprocessed following published methods with currently accepted mathematical models. This processed data is then projected against a databased model of pluripotency derived from the transcriptional profiles of hundreds of validated pluripotent and somatic cell lines. This projection allows for the direct comparison of our submitted sample data to hundreds of previously characterized data sets.

In TeratoScore In order to quantitatively estimate the pluripotency of the tumor initiating cells, we calculated the mean gene expression of each lineage and extraembryonic tissue, multiplied them, and divided by 1,000, producing a single score. This analysis, termed TeratoScore, essentially estimates the differentiation potency of the tumor initiating cells—the very goal of Teratoma formation.

3.) On the basis of Consideration on type of Microarray expression data:

The microarray data, which is available on NCBI's GEO Dataset, is present in different forms or we can say that the platforms that were used for cell lines are different for different cells depending upon the requirement of researcher. Different tools accept the files in their default platforms e.g., PluriTest accepts only ILLUMINA microarray data whereas CellNet and TeratoScore accept Affymetrix data.

4.) Statistical parameters :

These tools use different Statistical parameters viz. In CellNet several statistical parameters used for calculation of Gene regulatory network, it uses Z-score analysis and MEAN. Whereas PluriTest involves the use of Quantile Normalization, Loess Normalization (both for normalizing the variance stabilizing transformation values for the reduction in the spurious variations came during fluorescent dyes color combination analysis).

5.) Results Visualization:

The results we got from these analyses are in different forms for individual tool e.g. In CellNet the result window shows Heatmap and histograms based on the gene regulatory network analysis. In PluriTest the results consists of pluripotency graph, novelty graph, boxplots, hierarchical clustering and in TeratoScore the results include only the bargraph depicting the TeratoScore outcome. The Bars shows cells score (Teratomas forming score) with respect to that outcome.

6.) All the tools are performing using R script at Backend.

7.) Parameters Consideration:

PluriTest and CellNet use several parameters on which their working relies i.e. by considering these parameters they actually define the state of cell. These parameters include Telomerase activity, Differentiation, Methylation, Histone Demethylation etc. (as discussed earlier).Whereas TeratoScore paradigm based upon finding of level of Teratomas formation after the iPSCs reprogramming.

8. Rate of Analysis:

The most important thing in any online analysis is the time taken by that particular software. Here, we notice that the Rate of analysis is much faster and frequent in PluriTest as compared to others. In case of CellNet and TeratoScore, the results were sent through mail to the user's given account, but this process at least took 15mins to 2hrs depending upon the net connectivity.

9. Based on no. and size of the Microarray file to be uploaded:

File size of cell line and number of files included in the particular dataset chosen from NCBI is also a great concern in these tools. As for PluriTest, we can choose maximum of 12 files only of the cell lines present in the dataset. In CellNet, it is not mentioned for number of files but the total size of the cell lines to adopt is not more than 128MB. TeratoScore has no such obligations but the chosen file must be of Affymetrix Human Genome U133 plus 2.0 array platform.

10. Based on Portability, Accessibility and Reliability:

If we consider all the three tools we found that PluriTest found to be more Accessible and reliable with more accuracy in results, also it provides us space where our previous work could be saved for later view but in other two cases we don't have this facility online but we can review our results anytime by accessing through our mail IDs. If we talk about Security then CellNet and TeratoScore found to be more secured because our results are not saved online in tools provided database but we got them in our mail Ids which makes it more protective.

3) Development of a method to find pluripotency of any cell using TEXT file format:

- 3.1) First, we got the list of marker genes available in any stem cell whose presence decides the pluripotency level of any undifferentiated cell. These genes got isolated by using different techniques viz. MACS (Magnetic Cell Sorting Technique) and FCM (flow cytometry) technique which are one of the most effective cell isolating methods (**Source : Data collected from Embryonic Stem Cell Markers; Wenxiu Zhao, Xiang Ji, Fangfang Zhang, Liang Li and Lan Ma 2012)**

Table 1. ESC surface markers.

SSEAs markers	Characteristics	Classification
SSEA-1 (CD15/Lewis x)	Murine embryos, mouse ES cells, mouse and human germ cells, embryonal carcinoma (EC) cells	Carbohydrate-associated molecules
SSEA-3	Primate ES cells, human embryonic germ cells, human ES cells, embryonal carcinoma (EC) cells	Carbohydrate-associated molecules
SSEA-4	Primate ES cells, human embryonic germ cells, human ES cells, embryonal carcinoma (EC) cells	Carbohydrate-associated molecules
CD markers		
CD324 (E-Cadherin)	Human ES cells, mouse ES cells, embryonal carcinoma (EC) cells	Surface marker (Binding to integrin alphaE/beta7, homotypic interactions mediate cell adhesion)
CD90 (Thy-1)	Human ES cells, mouse ES cells, hematopoietic stem cells, embryonal carcinoma (EC) cells	Surface marker (hematopoietic stem cell and neuron differentiation, T activation)
CD117 (c-KIT, SCFR)	Human ES cells, mouse ES cells, hematopoietic stem progenitors, neural crest-derived melanocytes, primordial germ cells, embryonal carcinoma (EC) cells	Surface marker (Stem Cell Factor receptor)
CD326	Human ES cells, mouse ES cells, embryonal carcinoma (EC) cells	Surface marker (function as growth factor receptor or adhesion molecule)

Table 1. Cont.

SSEAs markers	Characteristics	Classification
CD9 (MRP1, TM4SF DRAP-27, p24)	Human ES cells, mouse ES cells	Surface marker (cell adhesion, migration, T co-stimulation)
CD29 ($\beta 1$ integrin)	Human ES cells, mouse ES cells	Surface marker
CD24 (HAS)	Human ES cells, mouse ES cells	Surface marker (T co-stimulation, CD62P receptor)
CD59 (Protectin)	Human ES cells, mouse ES cells	Surface marker (binds complement C8 and C9, blocks membrane attack complex assembly)
CD133	Human ES cells, mouse ES cells, embryonal carcinoma (EC) cells, Hematopoietic stem cells	Surface marker
CD31 (PECAM-1)	Human ES cells, mouse ES cells	Surface marker (CD38 receptor, signaling, platelet-endothelium adhesion)
CD49f (Integrin $\alpha 6$ /CD29)	Human ES cells, mouse ES cells	Membrane receptors

Markers		
TRA-1-60	Human ES cells, teratocarcinoma, embryonic germ cells, embryonal carcinoma (EC) cells	Surface antigen
TRA-1-81	Human ES cells, teratocarcinoma, embryonic germ cells, embryonal carcinoma (EC) cells	Surface antigen
Frizzled5	Human ES cells, mouse ES cells	Seven transmembrane-spanning G-protein-coupled receptor
Stem cell factor (SCF or c-Kit ligand)	ES cells, mouse ES cells, Hematopoietic stem cells, Mesenchymal stem cells, embryonal carcinoma (EC) cells	Cytokine, exist both as a transmembrane protein and a soluble protein
Cripto (TDGF-1)	Mouse ES cells, human ES cells, cardiomyocyte, embryonal carcinoma (EC) cells	Receptor for the TGF- β signaling pathway

Table 2. Transcription Factors.

CORE Nuclear transcription factors	Characteristics	Classification
Oct-3/4 (Pou5f1)	Mouse ES cells, human ES cells, embryonal carcinoma (EC) cells	POU family Transcription factors
Sox2	Mouse ES cells, human ES cells, embryonal carcinoma (EC) cells, neural stem (NS) cells	POU family binder Transcription factors
KLF4	Mouse ES cells, human ES cells, embryonal carcinoma (EC) cells	Zinc-finger Transcription factors
Nanog	Mouse ES cells, human ES cells, embryonal carcinoma (EC) cells	Transcription factors
Markers		
Rex1 (Zfp42)	Mouse ES cells, human ES cells, embryonal carcinoma (EC) cells	Zinc-finger Transcription factor
UTF1	Mouse human ES cells, germ line tissues in mouse and human, embryonal carcinoma (EC) cells	Transcriptional coactivator
ZFX	Murine ES cells, human ES cells, hematopoietic stem cells, embryonal carcinoma (EC) cells	X-linked zinc finger protein; Probable transcriptional activators
TBN	Mouse, human inner cell mass	New class of proteins with an important function in development
FoxD3	Murine ES cells, human ES cells, embryonal carcinoma (EC) cells	Forkhead Box family, transcriptional regulator
HMGA2	Mouse ES cells, human ES cells	Architectural transcription factors

NAC1	Mouse ES cells, human ES cells	The POZ/BTB domain family, nuclear factor
GCNF (NR6A1)	Mouse ES cells, human ES cells, embryonal carcinoma (EC) cells	Nuclear receptor gene superfamily, nuclear receptor
Stat3	Murine ES cells, Human ES cells, embryonal carcinoma (EC) cells	Transcription factor
LEF1, TCF3	Mouse ES cells, Human ES cells, embryonal carcinoma (EC) cells	(HMG) DNA binding protein family, Transcription factor
Sall4	Murine ES cells, Human ES cells, embryonal carcinoma (EC) cells	Zinc finger transcription factor
Fbxo15	Mouse ES cells, early embryos, and testis tissue, embryonal carcinoma (EC) cells	F-box protein family, target of Oct3/4
ECAT genes		
ECAT11 (FLJ10884/L1TD1)	Human ES cells, embryonal carcinoma (EC) cells	Downstream target of Nanog
Ecat1	Mouse oocytes, embryonal carcinoma (EC) cells	KH domain containing RNA binding protein
ECAT9 (Gdf3)	Human ES cells, embryonal carcinoma (EC) cells	TGF β superfamily, BMP inhibitor

Dppa genes		
Dppa5 (ESG1)	Mouse ES cells, Human ES cells, embryonal carcinoma (EC) cells	K homology RNA-binding (KH) domain
Dppa4	Mouse ES cells, Human ES cells, embryonal carcinoma (EC) cells	Nuclear factor
Dppa2 (ECSA)	Mouse ES cells, Human ES cells, embryonal carcinoma (EC) cells	DNA-binding protein
Dppa3 (Stella)	Mouse ES cells, human ES cells, embryonal carcinoma (EC) cells, primordial germ cells, oocytes, preimplantation embryos	Maternal factor

Table 3. Signal pathway-related intracellular markers.

Markers	Characteristics	Classification
SMAD1/5/8	Mouse ES cells, embryonal carcinoma (EC) cells	Smad proteins ((R-Smad), BMP signalling pathway
SMAD4	Mouse ES cells, human ES cells, embryonal carcinoma (EC) cells, early embryos, and testis tissue	Smad proteins (Co-SMAD), TGF- β /Activin/Nodal signalling pathway, BMP signalling pathway

Markers	Characteristics	Classification
SMAD2/3	Human ES cells, embryonal carcinoma (EC) cells	Smad proteins ((R-Smad), TGF- β /Activin/Nodal signaling pathway
β -catenin	Mouse ES cells, human ES cells, embryonal carcinoma (EC) cells	Transcription activators, Wnt/ β -catenin signaling pathway

CD9 (DRAP-27, MRP-1, p24)	Small cell lung cancer	Tetraspanin superfamily
CD59	Lung cancer, breast cancer	Membrane attack complex inhibition factor
CD326 (EpCAM)	Hepatocellular carcinoma	Wnt-B-catenin signaling target gene
Nanog	Brain and other kinds of cancer stem cells	Transcription factors
SOX2	Breast cancer stem cells, breast tumor	Transcription factors

Table 5. Tumor stem cell markers.

Overlapping with ESCs	Characteristics	Classification
SSEA-1 (CD15/ Lewis x)	Brain tumor stem cells	Surface marker
SSEA-3	Breast cancer stem cells, breast Cancer	Surface marker
SSEA-4	Breast Cancer, epithelial ovarian carcinoma	Surface marker
TRA-1-60	Germ cell tumor	Surface marker
CD133	Pancreatic exocrine cancer; colon cancer; glioma;	Surface marker
CD29	Mouse mammary cells	Surface marker
CD24	Breast tumor, tumor invasion, prostate cancer	Surface marker
CD90	Hepatocellular carcinoma cell lines	Surface marker

Table 5. Cont.

Overlapping with ESCs	Characteristics	Classification
OCT4	Bladder cancer, Lung Cancer stem cells	Transcription factors
KLF4	Breast cancer stem cells	Transcription factors
Rex1 (Zfp42)	Prostate Cancer, Renal cell carcinoma	Zinc-finger protein-42
UTF-1	Testicular germ cell tumours	Undifferentiated embryonic cell transcription factor 1
ZFX	Gastric cancer	Zinc-finger protein family
LEF1	Prostate cancer, lymphoblastic leukemia	Lymphoid enhancer-binding factor
SALL4	Breast cancer, leukemogenesis, testicular germ cell tumors	Transcription factor
ECAT9 (Gdf3)	Melanoma, Breast Carcinoma, Seminoma, germ cell tumors	TGF β superfamily, BMP inhibitor
DPPA2 (ECSA)	Lung Cancer	DNA-binding protein
HMGA2	Pancreatic adenocarcinoma, bladder cancer	Architectural transcription factors
NAC1	Breast, renal cell, and hepatocellular carcinoma	Nuclear factor
Cripto	Broad range of tumors	Extra-cellular plasma membrane growth factor
Stat3	A number of human tumor	Transcription factor
Tumor stem cell markers		
ALDH1	Lung tumor	Surface marker
Musashi-1	Endometrium tumor stem cells	Neural stem cell regulatory protein
LgR5	Esophageal adenocarcinomas	G-protein coupled receptor
DCAMKL-1	Intestinal neoplasia; adenoma stem cells	Related to β -catenin

TIM3	Myeloid leukemia (AML) stem cell	Surface marker
Brcal	Mammary tumor	Human caretaker gene
SDF-1, CXCR4	Homing of stem cells and metastasis of cancer cells	Cell factor
PSCA	Prostate cancer	Cell surface antigen
CD96	leukemic stem cells	Surface marker
CD44	Breast tumor, tumor invasion, prostate cancer	Surface marker
CD45	Hepatocellular carcinoma cell lines	Surface marker

Table 1: Showing 175 Marker genes which are present in Stem cells.

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

3.2) 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 assure 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.



Fig.1 Showing Interaction network between various Pluripotency Marker Genes using STRING.

Here in the previous network each sphere (node) represents particular gene. The STRING database quantify the uncertainty of interaction process by assigning scores to proposed protein interactions based on the nature and quality of the supporting evidence. STRING contains functional protein associations derived from in-house predictions and homology transfers, as well as taken from a number of externally maintained databases. Each of these interactions is assigned a score between zero and one, which is (meant to be) the probability that the interaction really exists given the available evidence. By using this entity we take out the nodes with highest scores and again made an interaction network for these selected nodes. By this we got 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.**

node1	node2	node1 accession	node2 accession	node1 annotation	node2 annotation	score
KLF4	EHMT1	ENSP00000363804	ENSP00000417980	Kruppel-like factor 4 (gut)	Euchromatic histone-lysine N-methyltran...	0.465
KLF4	EHMT2	ENSP00000363804	ENSP00000364687	Kruppel-like factor 4 (gut)	Euchromatic histone-lysine N-methyltran...	0.580
KLF4	EZH2	ENSP00000363804	ENSP00000320147	Kruppel-like factor 4 (gut)	Enhancer of zeste homolog 2 (Drosophil...	0.433
KLF4	FBXO15	ENSP00000363804	ENSP00000393154	Kruppel-like factor 4 (gut)	F-box protein 15; Substrate-recognition c...	0.686
KLF4	FUT4	ENSP00000363804	ENSP00000351602	Kruppel-like factor 4 (gut)	Fucosyltransferase 4 (alpha (1,3) fucosyl...	0.549
KLF4	GDF3	ENSP00000363804	ENSP00000331745	Kruppel-like factor 4 (gut)	Growth differentiation factor 3	0.525
KLF4	KIT	ENSP00000363804	ENSP00000288135	Kruppel-like factor 4 (gut)	V-kit Hardy-Zuckerman 4 feline sarcoma ...	0.498
KLF4	LGR5	ENSP00000363804	ENSP00000266674	Kruppel-like factor 4 (gut)	Leucine-rich repeat containing G protein...	0.652
KLF4	NANOG	ENSP00000363804	ENSP00000229307	Kruppel-like factor 4 (gut)	Nanog homeobox; Transcription regulato...	0.994
KLF4	PLAUR	ENSP00000363804	ENSP00000339328	Kruppel-like factor 4 (gut)	Plasminogen activator, urokinase recepto...	0.696
KLF4	POU5F1	ENSP00000363804	ENSP00000259915	Kruppel-like factor 4 (gut)	POU class 5 homeobox 1; Transcription f...	0.994
KLF4	PROM1	ENSP00000363804	ENSP00000415481	Kruppel-like factor 4 (gut)	Prominin 1; May play a role in cell differe...	0.504
KLF4	SALL4	ENSP00000363804	ENSP00000217086	Kruppel-like factor 4 (gut)	Sal-like 4 (Drosophila); Transcription fact...	0.919
KLF4	SMAD2	ENSP00000363804	ENSP00000262160	Kruppel-like factor 4 (gut)	SMAD family member 2; Receptor-regula...	0.938
KLF4	SMAD4	ENSP00000363804	ENSP00000341551	Kruppel-like factor 4 (gut)	SMAD family member 4; In muscle physi...	0.926
KLF4	SMARCA4	ENSP00000363804	ENSP00000350720	Kruppel-like factor 4 (gut)	SWI/SNF related, matrix associated, acti...	0.745
KLF4	SOX2	ENSP00000363804	ENSP00000323588	Kruppel-like factor 4 (gut)	SRY (sex determining region Y)-box 2; Tr...	0.993
KLF4	STAT3	ENSP00000363804	ENSP00000264657	Kruppel-like factor 4 (gut)	Signal transducer and activator of transc...	0.707
KLF4	SUZ12	ENSP00000363804	ENSP00000316578	Kruppel-like factor 4 (gut)	Suppressor of zeste 12 homolog (Drosop...	0.424
KLF4	TDGF1	ENSP00000363804	ENSP00000296145	Kruppel-like factor 4 (gut)	Teratocarcinoma-derived growth factor 1...	0.447

node1	node2	node1 accession	node2 accession	node1 annotation	node2 annotation	score
NANOG	POU5F1	ENSP00000229307	ENSP00000259915	Nanog homeobox; Transcription regulato...	POU class 5 homeobox 1; Transcription ...	0.999
NANOG	PROM1	ENSP00000229307	ENSP00000415481	Nanog homeobox; Transcription regulato...	Prominin 1; May play a role in cell differe...	0.737
NANOG	PTPRC	ENSP00000229307	ENSP00000356346	Nanog homeobox; Transcription regulato...	Protein tyrosine phosphatase, receptor t...	0.522
NANOG	SALL4	ENSP00000229307	ENSP00000217086	Nanog homeobox; Transcription regulato...	Sal-like 4 (Drosophila); Transcription fact...	0.992
NANOG	SMAD2	ENSP00000229307	ENSP00000262160	Nanog homeobox; Transcription regulato...	SMAD family member 2; Receptor-regula...	0.937
NANOG	SMAD4	ENSP00000229307	ENSP00000341551	Nanog homeobox; Transcription regulato...	SMAD family member 4; In muscle physi...	0.935
NANOG	SMARCA4	ENSP00000229307	ENSP00000350720	Nanog homeobox; Transcription regulato...	SWI/SNF related, matrix associated, acti...	0.637
NANOG	SOX2	ENSP00000229307	ENSP00000323588	Nanog homeobox; Transcription regulato...	SRY (sex determining region Y)-box 2; Tr...	0.999
NANOG	STAT3	ENSP00000229307	ENSP00000264657	Nanog homeobox; Transcription regulato...	Signal transducer and activator of transc...	0.977
NANOG	SUZ12	ENSP00000229307	ENSP00000316578	Nanog homeobox; Transcription regulato...	Suppressor of zeste 12 homolog (Droso...	0.596
NANOG	TDGF1	ENSP00000229307	ENSP00000296145	Nanog homeobox; Transcription regulato...	Teratocarcinoma-derived growth factor 1...	0.970
NANOG	THY1	ENSP00000229307	ENSP00000284240	Nanog homeobox; Transcription regulato...	Thy-1 cell surface antigen; May play a ro...	0.630
NANOG	UTF1	ENSP00000229307	ENSP00000305906	Nanog homeobox; Transcription regulato...	Undifferentiated embryonic cell transcrip...	0.672
NANOG	ZFX	ENSP00000229307	ENSP00000304985	Nanog homeobox; Transcription regulato...	Zinc finger protein, X-linked; Probable tra...	0.405
NCAM1	B3GAT1	ENSP00000318472	ENSP00000307875	Neural cell adhesion molecule 1; This pr...	Beta-1,3-glucuronyltransferase 1 (glucur...	0.824
NCAM1	CD34	ENSP00000318472	ENSP00000310036	Neural cell adhesion molecule 1; This pr...	CD34 molecule; Possible adhesion mole...	0.820
NCAM1	CD38	ENSP00000318472	ENSP00000226279	Neural cell adhesion molecule 1; This pr...	CD38 molecule; Synthesizes the second ...	0.736
NCAM1	CD44	ENSP00000318472	ENSP00000398632	Neural cell adhesion molecule 1; This pr...	CD44 molecule (Indian blood group); Re...	0.578
NCAM1	CD48	ENSP00000318472	ENSP00000357025	Neural cell adhesion molecule 1; This pr...	CD48 molecule; Ligand for CD2. Might fa...	0.747
NCAM1	CD58	ENSP00000318472	ENSP00000358501	Neural cell adhesion molecule 1; This pr...	CD58 molecule; Ligand of the T-lymphoc...	0.574

node1	node2	node1 accession	node2 accession	node1 annotation	node2 annotation	score
POU5F1	DNMT3L	ENSP00000259915	ENSP00000270172	POU class 5 homeobox 1; Transcription f...	DNA (cytosine-5-)-methyltransferase 3-lik...	0.457
POU5F1	DPPA2	ENSP00000259915	ENSP00000417710	POU class 5 homeobox 1; Transcription f...	Developmental pluripotency associated 2...	0.463
POU5F1	DPPA3	ENSP00000259915	ENSP00000339250	POU class 5 homeobox 1; Transcription f...	Developmental pluripotency associated 3...	0.583
POU5F1	DPPA4	ENSP00000259915	ENSP00000335306	POU class 5 homeobox 1; Transcription f...	Developmental pluripotency associated 4...	0.954
POU5F1	DPPA5	ENSP00000259915	ENSP00000359396	POU class 5 homeobox 1; Transcription f...	Developmental pluripotency associated 5...	0.414
POU5F1	EHMT2	ENSP00000259915	ENSP00000364687	POU class 5 homeobox 1; Transcription f...	Euchromatic histone-lysine N-methyltran...	0.531
POU5F1	EPCAM	ENSP00000259915	ENSP00000263735	POU class 5 homeobox 1; Transcription f...	Epithelial cell adhesion molecule; May ac...	0.694
POU5F1	EZH2	ENSP00000259915	ENSP00000320147	POU class 5 homeobox 1; Transcription f...	Enhancer of zeste homolog 2 (Drosophila...	0.589
POU5F1	FBXO15	ENSP00000259915	ENSP00000393154	POU class 5 homeobox 1; Transcription f...	F-box protein 15; Substrate-recognition c...	0.684
POU5F1	FOXO3	ENSP00000259915	ENSP00000360157	POU class 5 homeobox 1; Transcription f...	Forkhead box D3; Binds to the consensus...	0.963
POU5F1	FUT4	ENSP00000259915	ENSP00000351602	POU class 5 homeobox 1; Transcription f...	Fucosyltransferase 4 (alpha (1,3) fucosyl...	0.714
POU5F1	GDF3	ENSP00000259915	ENSP00000331745	POU class 5 homeobox 1; Transcription f...	Growth differentiation factor 3	0.636
POU5F1	ITGB1	ENSP00000259915	ENSP00000303351	POU class 5 homeobox 1; Transcription f...	Integrin, beta 1 (fibronectin receptor, beta...	0.551
POU5F1	KIT	ENSP00000259915	ENSP00000288135	POU class 5 homeobox 1; Transcription f...	V-kit Hardy-Zuckerman 4 feline sarcoma ...	0.667
POU5F1	KITLG	ENSP00000259915	ENSP00000328280	POU class 5 homeobox 1; Transcription f...	KIT ligand; Ligand for the receptor-type p...	0.493
POU5F1	KLF4	ENSP00000259915	ENSP00000363804	POU class 5 homeobox 1; Transcription f...	Kruppel-like factor 4 (gut)	0.994
POU5F1	NANOG	ENSP00000259915	ENSP00000229307	POU class 5 homeobox 1; Transcription f...	Nanog homeobox; Transcription regulato...	0.999
POU5F1	NR6A1	ENSP00000259915	ENSP00000420267	POU class 5 homeobox 1; Transcription f...	Nuclear receptor subfamily 6, group A, m...	0.593
POU5F1	NTSE	ENSP00000259915	ENSP00000257770	POU class 5 homeobox 1; Transcription f...	5'-nucleotidase, ecto (CD73); Hydrolyzes ...	0.513
POU5F1	PROM1	ENSP00000259915	ENSP00000415481	POU class 5 homeobox 1; Transcription f...	Prominin 1; May play a role in cell differe...	0.735

node1	node2	node1 accession	node2 accession	node1 annotation	node2 annotation	score
POU5F1	PTPRC	ENSP00000259915	ENSP00000356346	POU class 5 homeobox 1; Transcription ...	Protein tyrosine phosphatase, receptor t...	0.576
POU5F1	RNF2	ENSP00000259915	ENSP00000356480	POU class 5 homeobox 1; Transcription ...	Ring finger protein 2; E3 ubiquitin-protein...	0.483
POU5F1	SALL4	ENSP00000259915	ENSP00000217086	POU class 5 homeobox 1; Transcription ...	Sal-like 4 (Drosophila); Transcription fac...	0.986
POU5F1	SETDB1	ENSP00000259915	ENSP00000271640	POU class 5 homeobox 1; Transcription ...	SET domain, bifurcated 1; Histone meth...	0.756
POU5F1	SMAD2	ENSP00000259915	ENSP00000262160	POU class 5 homeobox 1; Transcription ...	SMAD family member 2; Receptor-regula...	0.939
POU5F1	SMAD3	ENSP00000259915	ENSP00000332973	POU class 5 homeobox 1; Transcription ...	SMAD family member 3; Receptor-regula...	0.477
POU5F1	SMAD4	ENSP00000259915	ENSP00000341551	POU class 5 homeobox 1; Transcription ...	SMAD family member 4; In muscle phys...	0.928
POU5F1	SMARCA4	ENSP00000259915	ENSP00000350720	POU class 5 homeobox 1; Transcription ...	SWI/SNF related, matrix associated, acti...	0.420
POU5F1	SOX2	ENSP00000259915	ENSP00000323588	POU class 5 homeobox 1; Transcription ...	SRY (sex determining region Y)-box 2; Tr...	0.998
POU5F1	STAT3	ENSP00000259915	ENSP00000264657	POU class 5 homeobox 1; Transcription ...	Signal transducer and activator of transc...	0.970
POU5F1	SUZ12	ENSP00000259915	ENSP00000316578	POU class 5 homeobox 1; Transcription ...	Suppressor of zeste 12 homolog (Droso...	0.571
POU5F1	TDGF1	ENSP00000259915	ENSP00000296145	POU class 5 homeobox 1; Transcription ...	Teratocarcinoma-derived growth factor ...	0.956
POU5F1	THY1	ENSP00000259915	ENSP00000284240	POU class 5 homeobox 1; Transcription ...	Thy-1 cell surface antigen; May play a ro...	0.658
POU5F1	UTF1	ENSP00000259915	ENSP00000305906	POU class 5 homeobox 1; Transcription ...	Undifferentiated embryonic cell transcri...	0.614
PROM1	ALDH1A1	ENSP00000415481	ENSP00000297785	Prominin 1; May play a role in cell differe...	Aldehyde dehydrogenase 1 family, mem...	0.702
PROM1	CD34	ENSP00000415481	ENSP00000310036	Prominin 1; May play a role in cell differe...	CD34 molecule; Possible adhesion mole...	0.881
PROM1	CD38	ENSP00000415481	ENSP00000226279	Prominin 1; May play a role in cell differe...	CD38 molecule; Synthesizes the second...	0.505
PROM1	CD44	ENSP00000415481	ENSP00000398632	Prominin 1; May play a role in cell differe...	CD44 molecule (Indian blood group); Re...	0.821
PROM1	CDH1	ENSP00000415481	ENSP00000261769	Prominin 1; May play a role in cell differe...	Cadherin 1, type 1, E-cadherin (epithelial...	0.578
PROM1	CXCL12	ENSP00000415481	ENSP00000379140	Prominin 1; May play a role in cell differe...	Chemokine (C-X-C motif) ligand 12; Che...	0.660

node1	node2	node1 accession	node2 accession	node1 annotation	node2 annotation	score
SOX2	NCAM1	ENSP00000323588	ENSP00000318472	SRY (sex determining region Y)-box 2; Tr...	Neural cell adhesion molecule 1; This pro...	0.463
SOX2	NTSE	ENSP00000323588	ENSP00000257770	SRY (sex determining region Y)-box 2; Tr...	5'-nucleotidase, ecto (CD73); Hydrolyzes ...	0.441
SOX2	POU5F1	ENSP00000323588	ENSP00000259915	SRY (sex determining region Y)-box 2; Tr...	POU class 5 homeobox 1; Transcription f...	0.998
SOX2	PROM1	ENSP00000323588	ENSP00000415481	SRY (sex determining region Y)-box 2; Tr...	Prominin 1; May play a role in cell differe...	0.735
SOX2	PTPRC	ENSP00000323588	ENSP00000356346	SRY (sex determining region Y)-box 2; Tr...	Protein tyrosine phosphatase, receptor ty...	0.491
SOX2	SALL4	ENSP00000323588	ENSP00000217086	SRY (sex determining region Y)-box 2; Tr...	Sal-like 4 (Drosophila); Transcription fact...	0.966
SOX2	SMAD2	ENSP00000323588	ENSP00000262160	SRY (sex determining region Y)-box 2; Tr...	SMAD family member 2; Receptor-regula...	0.942
SOX2	SMAD4	ENSP00000323588	ENSP00000341551	SRY (sex determining region Y)-box 2; Tr...	SMAD family member 4; In muscle physi...	0.936
SOX2	SMARCC1	ENSP00000323588	ENSP00000254480	SRY (sex determining region Y)-box 2; Tr...	SWI/SNF related, matrix associated, acti...	0.492

Now, we plot another network using the resultant common genes which we got from the parent network and it includes POU5F1, SOX2, NANOG, KLF4, SMAD2, SMAD4, SALL4 and DPPA4.

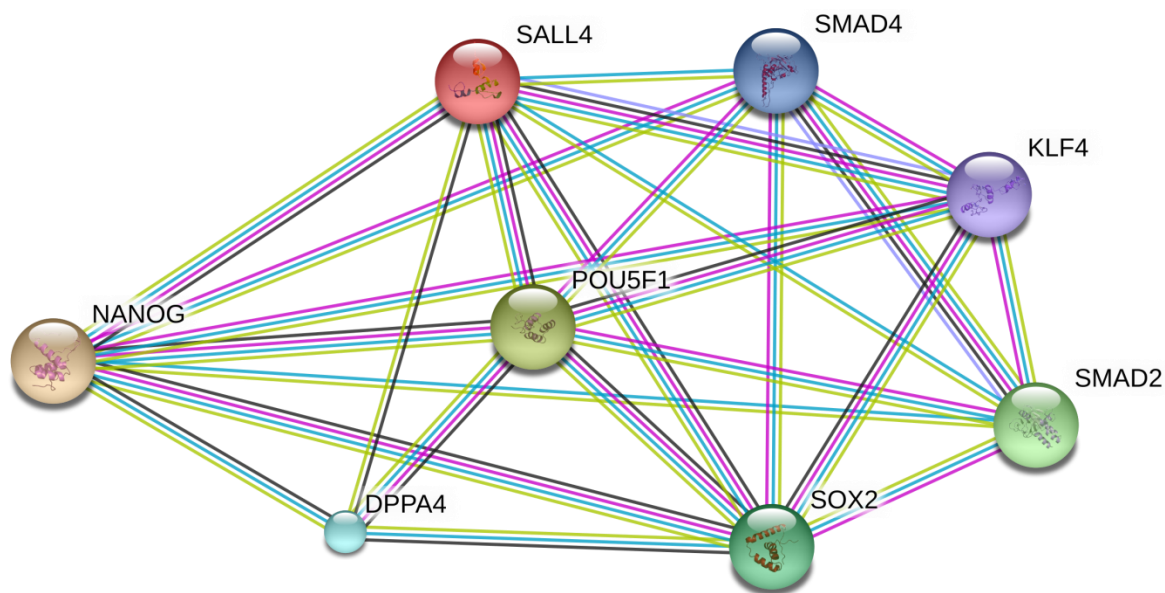


Fig.2 STRING network showing important genes responsible for pluripotency of any cell.

node1	node2	node1 accession	node2 accession	node1 annotation	node2 annotation	score
DPPA4	NANOG	ENSP00000335306	ENSP00000229307	Developmental pluripotency associated 4...	Nanog homeobox; Transcription regulato...	0.977
DPPA4	POU5F1	ENSP00000335306	ENSP00000259915	Developmental pluripotency associated 4...	POU class 5 homeobox 1; Transcription f...	0.954
DPPA4	SALL4	ENSP00000335306	ENSP00000217086	Developmental pluripotency associated 4...	Sal-like 4 (Drosophila); Transcription fact...	0.594
DPPA4	SOX2	ENSP00000335306	ENSP00000323588	Developmental pluripotency associated 4...	SRY (sex determining region Y)-box 2; Tra...	0.955
KLF4	NANOG	ENSP00000363804	ENSP00000229307	Kruppel-like factor 4 (gut)	Nanog homeobox; Transcription regulato...	0.994
KLF4	POU5F1	ENSP00000363804	ENSP00000259915	Kruppel-like factor 4 (gut)	POU class 5 homeobox 1; Transcription f...	0.994
KLF4	SALL4	ENSP00000363804	ENSP00000217086	Kruppel-like factor 4 (gut)	Sal-like 4 (Drosophila); Transcription fact...	0.919
KLF4	SMAD2	ENSP00000363804	ENSP00000262160	Kruppel-like factor 4 (gut)	SMAD family member 2; Receptor-regulat...	0.938
KLF4	SMAD4	ENSP00000363804	ENSP00000341551	Kruppel-like factor 4 (gut)	SMAD family member 4; In muscle physio...	0.926
KLF4	SOX2	ENSP00000363804	ENSP00000323588	Kruppel-like factor 4 (gut)	SRY (sex determining region Y)-box 2; Tra...	0.993
NANOG	DPPA4	ENSP00000229307	ENSP00000335306	Nanog homeobox; Transcription regulato...	Developmental pluripotency associated 4...	0.977
NANOG	KLF4	ENSP00000229307	ENSP00000363804	Nanog homeobox; Transcription regulato...	Kruppel-like factor 4 (gut)	0.994
NANOG	POU5F1	ENSP00000229307	ENSP00000259915	Nanog homeobox; Transcription regulato...	POU class 5 homeobox 1; Transcription f...	0.999
NANOG	SALL4	ENSP00000229307	ENSP00000217086	Nanog homeobox; Transcription regulato...	Sal-like 4 (Drosophila); Transcription fact...	0.992
NANOG	SMAD2	ENSP00000229307	ENSP00000262160	Nanog homeobox; Transcription regulato...	SMAD family member 2; Receptor-regulat...	0.937
NANOG	SMAD4	ENSP00000229307	ENSP00000341551	Nanog homeobox; Transcription regulato...	SMAD family member 4; In muscle physio...	0.935
NANOG	SOX2	ENSP00000229307	ENSP00000323588	Nanog homeobox; Transcription regulato...	SRY (sex determining region Y)-box 2; Tra...	0.999
POU5F1	DPPA4	ENSP00000259915	ENSP00000335306	POU class 5 homeobox 1; Transcription f...	Developmental pluripotency associated 4...	0.954
POU5F1	KLF4	ENSP00000259915	ENSP00000363804	POU class 5 homeobox 1; Transcription f...	Kruppel-like factor 4 (gut)	0.994
POU5F1	NANOG	ENSP00000259915	ENSP00000229307	POU class 5 homeobox 1; Transcription f...	Nanog homeobox; Transcription regulato...	0.999

From this table too, we take out most frequent and highest scorer Nodes. We found POU5F1, SOX2, NANOG, KLF4. Now, we again make our interaction network using STRING which is consist of these 4 genes.

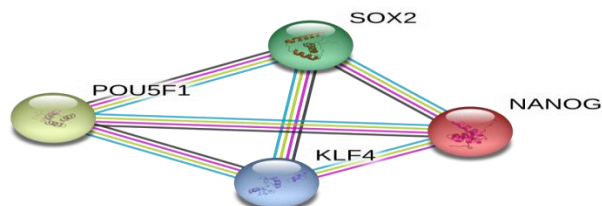


Fig.3 Showing final network

node1	node2	node1 accession	node2 accession	node1 annotation	node2 annotation	score
KLF4	NANOG	ENSP00000363804	ENSP00000229307	Kruppel-like factor 4 (gut)	Nanog homeobox; Transcription regulato...	0.994
KLF4	POU5F1	ENSP00000363804	ENSP00000259915	Kruppel-like factor 4 (gut)	POU class 5 homeobox 1; Transcription f...	0.994
KLF4	SOX2	ENSP00000363804	ENSP00000323588	Kruppel-like factor 4 (gut)	SRY (sex determining region Y)-box 2; Tra...	0.993
NANOG	KLF4	ENSP00000229307	ENSP00000363804	Nanog homeobox; Transcription regulato...	Kruppel-like factor 4 (gut)	0.994
NANOG	POU5F1	ENSP00000229307	ENSP00000259915	Nanog homeobox; Transcription regulato...	POU class 5 homeobox 1; Transcription f...	0.999
NANOG	SOX2	ENSP00000229307	ENSP00000323588	Nanog homeobox; Transcription regulato...	SRY (sex determining region Y)-box 2; Tra...	0.999
POU5F1	KLF4	ENSP00000259915	ENSP00000363804	POU class 5 homeobox 1; Transcription f...	Kruppel-like factor 4 (gut)	0.994
POU5F1	NANOG	ENSP00000259915	ENSP00000229307	POU class 5 homeobox 1; Transcription f...	Nanog homeobox; Transcription regulato...	0.999
POU5F1	SOX2	ENSP00000259915	ENSP00000323588	POU class 5 homeobox 1; Transcription f...	SRY (sex determining region Y)-box 2; Tra...	0.998
SOX2	KLF4	ENSP00000323588	ENSP00000363804	SRY (sex determining region Y)-box 2; Tra...	Kruppel-like factor 4 (gut)	0.993
SOX2	NANOG	ENSP00000323588	ENSP00000229307	SRY (sex determining region Y)-box 2; Tra...	Nanog homeobox; Transcription regulato...	0.999
SOX2	POU5F1	ENSP00000323588	ENSP00000259915	SRY (sex determining region Y)-box 2; Tra...	POU class 5 homeobox 1; Transcription f...	0.998

From this, we finally got our genes which are having the most number of interaction and with highest no. of scores. 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.

3.3) 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 frequent availability.

Method for determining Pluripotency by using TEXT file format:

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.

(Note: The Data to be arranged separately for each cell type i.e. for ESC, IPSC, Somatic etc.)

	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

Fig.4 Showing Non-Normalized Microarray data (GSE72078) in excel sheet collected from GEO datasets.

	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

Fig.5 Showing Normalized microarray data (using Simple normalization function).

$$\rightarrow \text{Mean} = \sum (x_i) / n_i$$

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

After this we take our 175 Marker genes. Now compare and match them with the existing list of NCBI downloaded expression dataset's gene list by using JAVA program (specially designed for finding of no. of marker genes present in the microarray data.

Table 2 : Gene List after Running Java Program:

BRCA1	CD68	CD96	EED	NANOG	SMARCA4
CD24	CD69	CXCR4	EHMT1	NR6A1	SMARCA5
CD34	CD70	CXXC1	EHMT2	POU5F1	SMARCB1
CD38	CD72	DICER1	EZH2	PSCA	SMARCC1
CD44	CD74	DNMT1	FBXO15	SALL4	SUZ12
CD47	CD80	DNMT3L	FOXD3	SETDB1	TCF3
CD48	CD81	DPPA2	GDF3	SMAD1	UTF1
CD52	CD82	DPPA3	HMGA2	SMAD2	ZFX
CD53	CD83	DPPA4	KLF4	SMAD3	
CD55	CD84	DPPA5	L1TD1	SMAD4	
CD58	CD86	SOX2	LEF1	SMAD5	
CD59	CD9	STAT3	SUV39H2	SALL4	

Total 68 Marker Genes are matched with the Genes of Microarray Data of GSE72078 cell lines.

	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	GSM1854270	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	1.00663113	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

Fig.6 Table showing marker genes with reference IDs and their respective expression values in each cell line. Also, Calculated Mean and log FC values are shown.

3.4) After normalization we took the data for prediction of Threshold using Quantile function through R Studio, this approach gives 5 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 said to be pluripotent.

Quantile function => $Q(p) = (1 - f) Q(p_i) + f Q(p_{i+1})$

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

Fig.7 Calculated Threshold value for detection of pluripotency using Quantile Normalization method through R script (Threshold = 0.6193309).

```
library(limma)

library (readxl)

n <- read_excel("C:/Users/my/Desktop/rt2.xlsx")

View(n)

k4$ID <- NULL

k4$`Gene Title` <- NULL

View(k4)

n2 <- n$logFC

quantile(n2,probs = c(0,0.25,0.5,0.75,1))
```

R SCRIPT

- 3.5) 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.

For TEST Samples:

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.

1	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

Fig.8 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).

Table 3: Gene List after Running Java Program:

BRCA1	CD59	CD86	EED	NANOG	SMARCA4
CD24	CD63	CD9	EHMT1	NR6A1	SMARCA5
CD34	CD68	CD96	EHMT2	POU5F1	SMARCB1
CD38	CD69	CXCR4	EZH2	PSCA	SMARCC1
CD44	CD70	CXXC1	FBXO15	PSCA	SOX2
CD46	CD72	DICER1	FOXD3	SALL4	STAT3
CD47	CD74	DNMT1	GDF3	SETDB1	SUV39H2
CD48	CD80	DNMT3L	HMGA2	SMAD1	SUZ12
CD52	CD81	DPPA2	KLF4	SMAD2	TCF3
CD53	CD82	DPPA3	L1TD1	SMAD3	UTF1
CD55	CD83	DPPA4	LEF1	SMAD4	ZFX
CD58	CD84	DPPA5	MBD3	SMAD5	

Total 71 Marker Genes Are Matched with the Genes of Microarray Data of GSE93228 cell lines

→ Test Sample Data:

1	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

Fig.9 Table after Matching our Marker genes within the data of test sample we got 71 matched entries by using JAVA developed program.

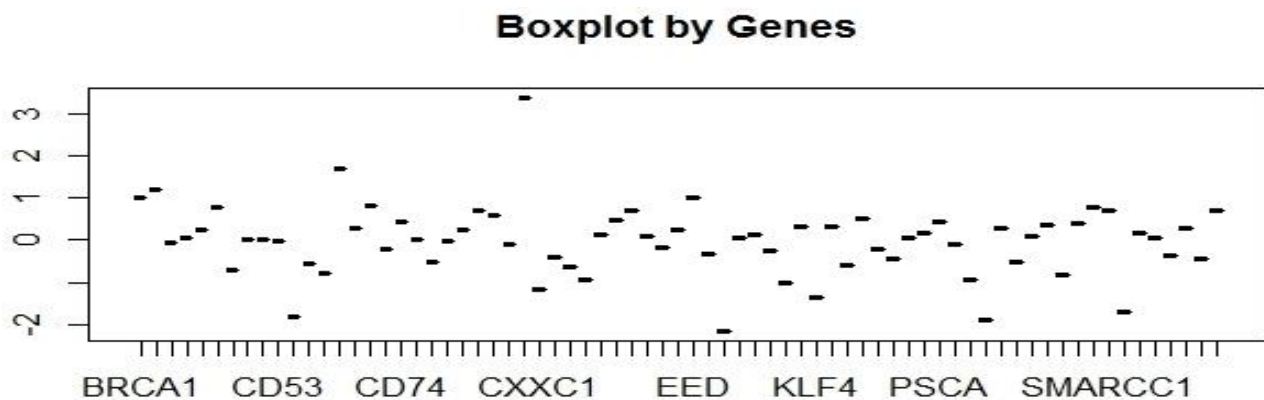


Fig.10 Boxplot showing non-normalized gene expression data of GSE93228, depicting discreteness of the values.

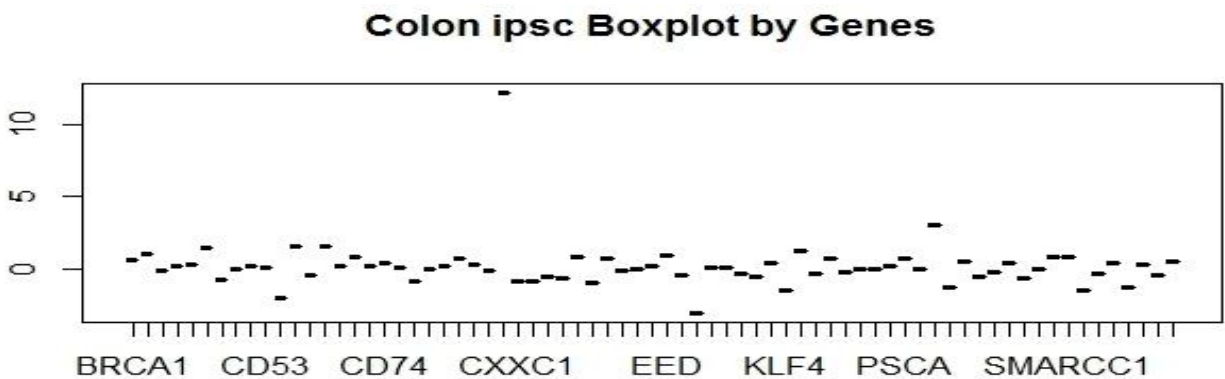


Fig.11 Boxplot showing Normalized gene expression values in a linear manner.


```

library(limma)

library(readxl)

n <- read_excel("C:/Users/my/Desktop/rt2.xlsx")

View(n)

names(n)

k4$ID <- NULL

k4$`Gene Title` <- NULL

View(k4)

n2 <- n$logFC

boxplot(n2)

quantile(n2, probs = c(0,0.25,0.5,0.75,1))

boxplot(n2, main = "Boxplot", ylab = "Nor.MEAN", las = 1)

boxplot(n2 ~ n3, main = 'Colon ipsc Boxplot by Genes')

```

R Script for
Boxplot

4							
5							Threshold
6							
7	0%	25%	50%	75%	100%		
8	-3.0712	-0.47081	0.046304	0.496912	12.16558		1.8333558
9							

Fig.12 The Pluripotency score is calculated again using R Script as earlier and it is passing the Threshold value. Hence we can say that the Cell line for Microarray data of GSE 93228 sample is pluripotent.

After this, we had randomly taken mast cell data just to check the accuracy of our pluripotency status detecting method. We repeat all the above steps as we done for our pass sample and the results are as under :

	A	B	C	D	E	F	G	H	I	J	K	L
1	"ID" "adj.P.Val"	Gene Symbol	GSM1349722	GSM1349723	GSM1349724	GSM1349728	GSM1349729	GSM1349730	MEAN	MEDIAN(Pos)	GSM1349725	GSM1349726
2	"11536" "0.000247"	EEF1A1	34223.62	32865.53	34223.62	34223.62	34223.62	34223.62	33997.27167	34223.62	32079.17	34223.62
3	"17777" "0.000247"	SLC35E2	15.81119	10.78991	14.33032	16.02723	20.04045	8.881108	14.313368	15.070755	19.98077	13.73935
4	"15425" "0.000247"	RPS28	2251.751	2091.537	4370	1863.689	4206.793	3944.467	3121.372833	3098.109	1676.106	3535.883
5	"816" "0.000247"	IPO13	26.83319	41.6515	25.39379	36.98999	37.05711	36.299	34.03743	36.644495	49.14602	50.87771
6	"631" "0.000247"	AFAP1	53.82867	47.67378	23.88739	45.82653	20.00691	32.83493	37.343035	39.33073	3.614952	4.090782
7	"3137" "0.000265"	0	16.25593	11.1529	9.570271	7.773258	4.830535	9.36933	9.825370667	9.4698005	13.48144	-0.6609073

	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.13 After the calculation of pluripotency score using Quantile function we found that the sample failed to pass the Threshold and hence we termed it as NON-Pluripotent cell sample.

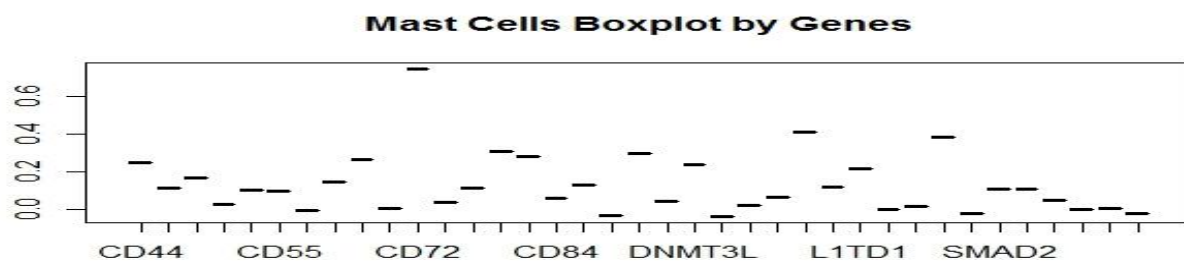


Fig.14 Boxplot showing gene expression values for Mast cells in a discrete manner after normalization.

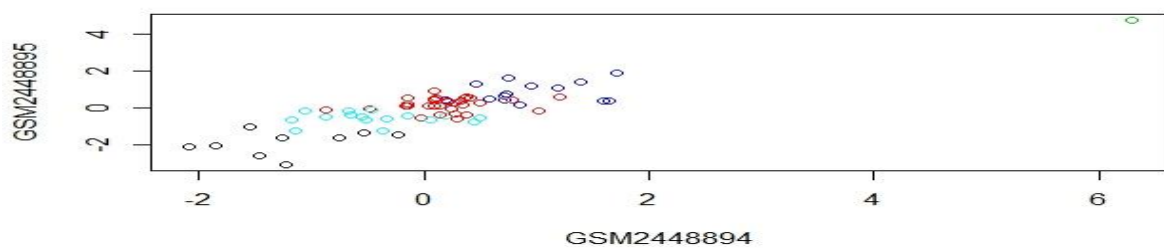


Fig.15 Graph showing clustered cell line from test sample on the basis of Gene expression data by using K-Means clustering through R Script for the development of Hierarchical clustering in Heatmap.

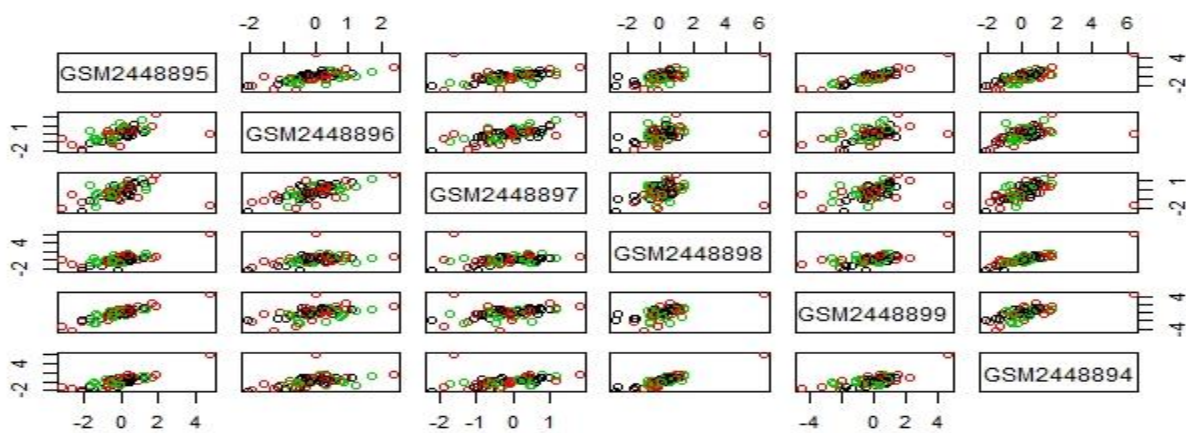


Fig.16 Figure showing K-Means clustered data of all 6 cell lines of test data sets created using R Script.

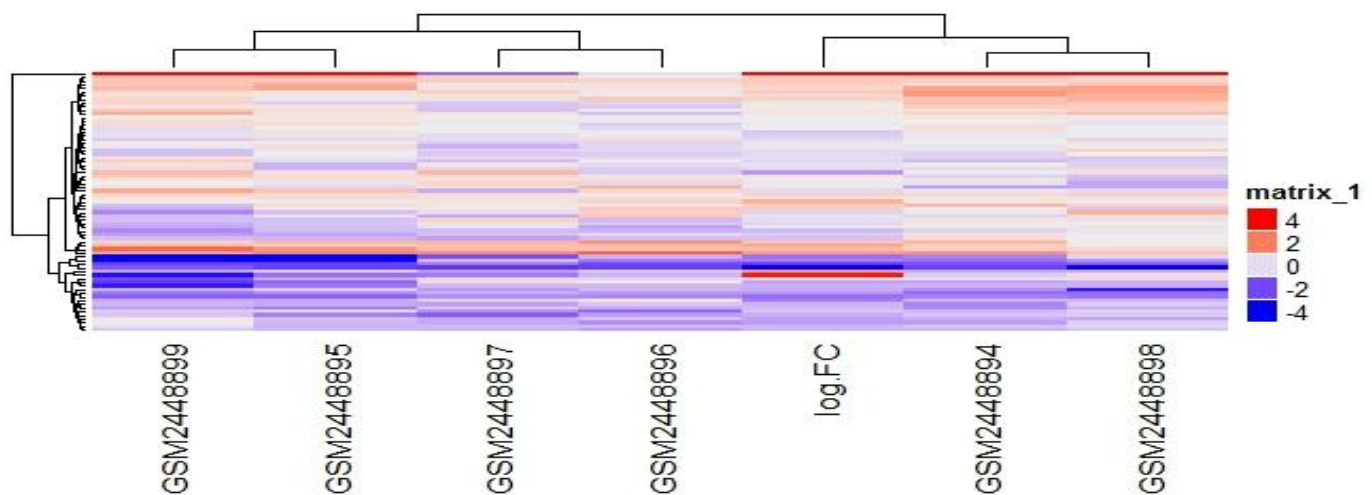


Fig.17 Heatmap Generated for all test sample's cell lines gene expression data using R program.

R Script for creating Heatmap using R Studio.

```
library(ComplexHeatmap)

library(grid)

library(stats)

filename1 <- "c:/Users/my/Desktop/new.txt"

k <- read.table(filename1, sep = "\t",quote = "",stringsAsFactors = FALSE, header = TRUE)

View(k)

#k$ID_REF__1 <- NULL

k3.features <- k

k3.features$MATCHED.GENES<-NULL

View(k3.features)

k$MEAN <- NULL

k$S.D. <- NULL

k$`MATCHED GENES` <- NULL

k3.features <- k

results <- kmeans(k3.features,3)

results

results$size

results$cluster

results$totss

results$centers

results$withinss

results$tot.withinss

results$betweenss

results$iter
```

```

results$ifault

table(k$MATCHED.GENES, results$cluster)

plot(k[c("GSM2448894","GSM2448895")],col = results$cluster)

plot(k[c("GSM2448895","GSM2448896","GSM2448897","GSM2448898","GSM2448899","GSM2448894"
)],col = 3:2)

my_matrix1 <- as.matrix(k3.features[1:6],c[])

class(results$cluster)

class(my_matrix1)

head(my_matrix1)

gene_info1 <- data.frame(gene = k3.features$MATCHED.GENES[200:400])

gene_info1

Heatmap(my_matrix1)

```

6) 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.

	IPSC Range		ESC Range	
	Is Data Manually Normalized or Not			
	YES	NO	YES	NO
A)	5.0 to 9.0	-1.5 to 5.0	2.0 to 3.0	-0.25 to 2.0
B)	2.0 to 4.9	-2.0 to -1.51	1.5 to 1.9	-3.0 to -0.249
C)	-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.

Result for Pluripotency level identification through JAVA developed program

➔ 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: GSM1854266

NANOG= 1.0564212

POU5F1= 0.88678443

SOX2= 0.42901477

Less Pluripotent Cell= GSM1854266

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: GSM1854267

NANOG= 0.668517

POU5F1= 0.81713855

SOX2= 1.1758825

Less Pluripotent Cell= GSM1854267

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: GSM1854268

NANOG= 0.73580366

POU5F1= 0.9113936

SOX2= 0.42615175

Less Pluripotent Cell= GSM1854268

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: GSM1854269

NANOG= 1.5778602

POU5F1= 1.2110461

SOX2= 0.12783645

Less Pluripotent Cell= GSM1854269

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: GSM1854270

NANOG= 1.45287005

POU5F1= 1.2110461

SOX2= 1.522041676

Less Pluripotent Cell= GSM1854270

➔ 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

For Pluripotency level determination we first took the control sample. Our program first check 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 consider the range for this condition and gave us the results that in which category or level the given sample is lying. As, we can see that our control sample results are least pluripotent in normalized IPSC range.

Same for 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 text format and .idat format) are shown in figures.

RESULT VALIDATION:

1) Text file based method:

GSE92706 (Differentiation of Human IPSC to mammary like organoids)

Range: ≥ 0.3 to ≤ 3.0

Range: GSM2435593

NANOG= 1.2516

POU5F1= 1.0151

SOX2= 1.1078

Highly Pluripotent Cell= GSM2435593

Range: ≥ 0.3 to ≤ 3.0

Range: GSM2435596

NANOG= 0.7761

POU5F1= 1.0205

SOX2= 1.0584

Highly Pluripotent Cell= GSM2435596

Range: ≥ 0.3 to ≤ 3.0

Range: GSM2435594

NANOG= 1.3013

POU5F1= 0.95

SOX2= 1.1776

Highly Pluripotent Cell= GSM2435594

Range: ≥ 0.3 to ≤ 3.0

Range: GSM2435597

NANOG= 0.9877

POU5F1= 0.997

SOX2= 0.8913

Highly Pluripotent Cell= GSM2435597

Range: ≥ 0.3 to ≤ 3.0

Range: GSM2435595

NANOG= 0.8062

POU5F1= 0.9371

SOX2= 0.9506

Highly Pluripotent Cell= GSM2435595

Range: ≥ 0.3 to ≤ 3.0

Range: GSM2435598

NANOG= 0.8768

POU5F1= 1.08

SOX2= 0.814

Highly Pluripotent Cell= GSM2435598

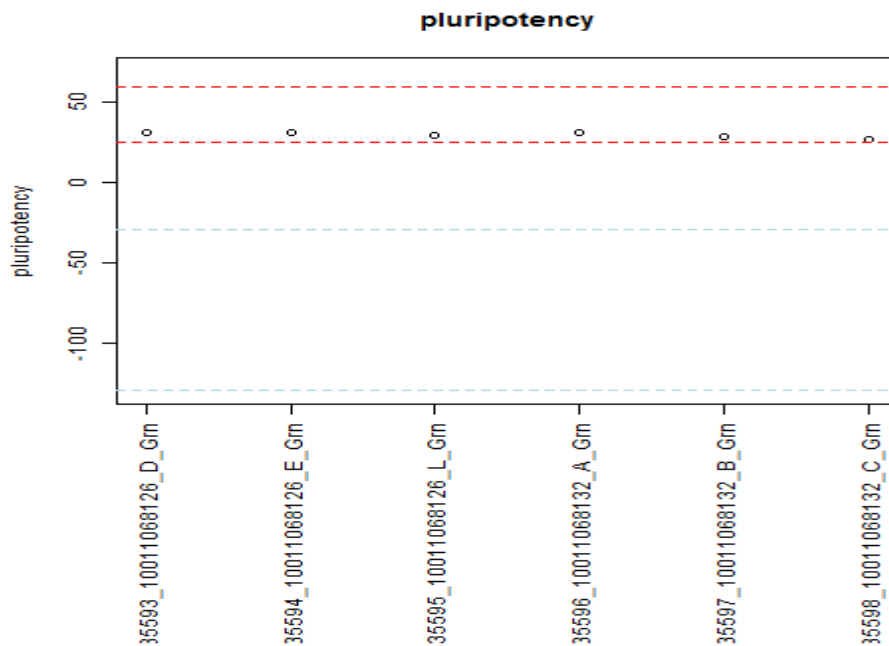
All, the cell lines are showing high pluripotency under the range set for manually Normalized data of IPSC category.

2) PLURITEST results for GSE92706

Table with results

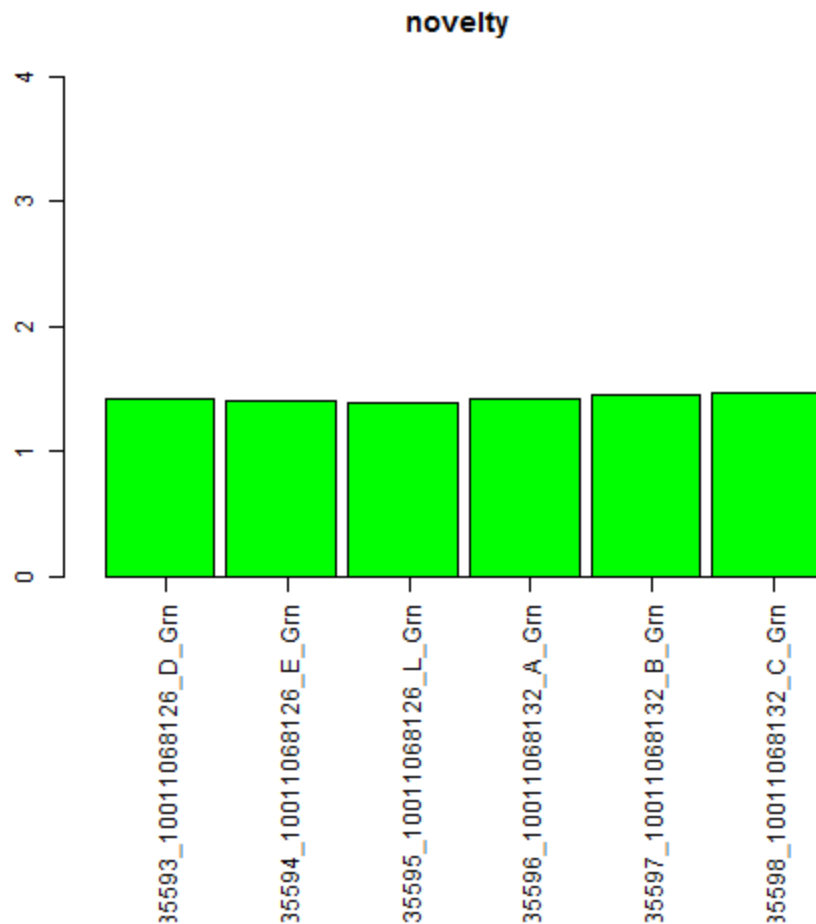
	pluri-raw	pluri logit-p	novelty	novelty logit-p	RMSD
GSM2435593_10011068126_D_Grn	31.55	1.00	1.42	0.01	0.46
GSM2435594_10011068126_E_Grn	31.52	1.00	1.41	0.01	0.46
GSM2435595_10011068126_L_Grn	29.95	1.00	1.38	0.01	0.44
GSM2435596_10011068132_A_Grn	31.11	1.00	1.42	0.01	0.46
GSM2435597_10011068132_B_Grn	28.55	1.00	1.46	0.02	0.46
GSM2435598_10011068132_C_Grn	26.88	1.00	1.46	0.02	0.47

Model-Based Multi-Class Pluripotency Score



The pluripotency graph is based on the pluri raw values present in the table above. The area between the red lines indicates the range that contains approx. 95 percent of the pluripotent samples tested. The pluripotency Score gives an indication if a sample contains a pluripotent signature. The Blue lines indicate those scores that we have observed in approx. 95 percent of the non-pluripotent samples.

Novelty



The novelty graph is based on the novelty score as given in the above table. The Novelty Score that is based on well-characterized pluripotent samples in the stem cell model matrix. Samples are color-coded green (pluripotent), orange, red (not-pluripotent) based on the probabilities given from the logistic regression model. Orange and red samples are more dissimilar to the pluripotent samples in the model matrix than the other pluripotent samples in the matrix.

These results are indicating us that both the approaches are providing favorable results, both the results are showing that the given cell line is Highly Pluripotent. Hence, we can now say that our text based method is also providing genuine results.

DISCUSSION:

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. As those tools is still dealing with some bugs as by considering 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 text file based logic of creating and arranging raw text file into matched and arranged file format with respect to 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 were 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 R script. This approach will surely provide an open access for identifying pluripotency and understanding the working and expression nature of various genes involved in reprogramming strategies.

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 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 and then after 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 with various problems associated with pre-existing online tools. We can make our own 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 are able to get Pluripotency state of any type of cell regardless of any particular platform or any file format.

FUTURE PERSPECTIVES:

1. Researches which are based on pluripotency state of cells have to gather info about the pluripotent state of that cell. This approach could be helpful for them.
2. By collecting data from various other databases, a tool can be created which could be more helpful in making process more reliable and frequent.
3. We can estimate easily which gene is playing a key role in determining the potency of that particular cell and is responsible for maintaining pluripotency in a cell.
4. By this approach researchers will be able to produce more number of iPSC's with enhanced efficiencies.
5. For deep and thorough study we have to consider various other ranges from different other genes.

REFERENCES

Carolina Perez-Iratxeta, Miguel A. Andrade-Navarro and Jonathan D.Wren. Evolving research trends in bioinformatics. Briefings in Bioinformatics. Oct 31, 2006. Vol 8. NO 2. 88.

Franz-Josef Müller^{1,11}, Bernhard M Schuldt^{2,11}, Roy Williams³, Dylan Mason⁴, Gulsah Altun⁵, Eirini P Papapetrou⁶, Sandra Danner⁷, Johanna E Goldmann^{5,8}, Arne Herbst¹, Nils O Schmidt⁹, Josef B Aldenhoff¹, Louise C Laurent^{5,10} & Jeanne F Loring⁵ A bioinformatic assay for pluripotency in human cells.

Gokhale PJ, Andrews PW: The development of pluripotent stem cells. *Curr Opin Genet Dev* 2012, 22:403–408.

Hitoshi Niwa, How is pluripotency determined and maintained? *Development* 134, 635-646 (2007) doi:10.1242/dev.02787.

Jing Hua Zhao and Qihua Tan “Integrated Analysis of Genetic Data with R” 2006 Human Genomics, vol 2, pp.258-265.

Jing Hua Zhao^{1*} and Qihua Tan² Integrated analysis of genetic data with R.

Loh, Y. H., Wu, Q., Chew, J. L., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J. et al. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* 38, 431-440.

Michael W Nestor and Scott A Noggle* Standardization of human stem cell pluripotency using bioinformatics Nestor and Noggle *Stem Cell Research & Therapy* 2013, 4:37.

Patrick Cahan^{1,#}, Hu Li^{2,#}, Samantha A. Morris^{1,#}, Edroaldo Lummertz da Rocha^{3,4}, George Q. Daley^{1,*}, and James J. Collins^{3,*} CellNet: Network Biology Applied to Stem Cell Engineering *Cell*. 2014 August 14; 158(4): 903–915. doi:10.1016/j.cell.2014.07.020.

Polani B.Ramesh Babu and P.Krishnamoorthy Applications of Bioinformatics Tools in Stem Cell Research: An Update *Journal of Pharmacy Research* 2012,5(9), 4863-4866.

Sandrine Dudoit ,Yee Hwa Yang, Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data.

Som A, Harder C, Greber B, Siatkowski M, Paudel Y, et al. (2010) The PluriNetWork: An Electronic Representation of the Network Underlying Pluripotency in Mouse, and Its Applications. *PLoS ONE* 5(12): e15165. doi:10.1371/journal.pone.0015165.

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, 131:861–872.

Ulf Tiemann,¹ Adele Gabriele Marthaler,¹ Kenjiro Adachi,¹ Guangming Wu,¹ Gerrit Ulf Lennart Fishedick,¹ Marcos Jesu´s Arau´zo-Bravo,¹ Hans Robert Scho´ler,^{1,*} and Natalia Tapia^{1,*} Counteracting Activities of OCT4 and KLF4 during Reprogramming to Pluripotency. <http://dx.doi.org/10.1016/j.stemcr.2014.01.005>

Vimal K.Singh^{1*}, Manisha Kalsan², Neeraj Kumar², Abhishek Saini² and Ramesh Chandra³ Induced pluripotent stem cells :applications in regenerative medicine, disease modeling, and drug discovery doi: 10.3389/fcell.2015.00002

Wenxiu Zhao ¹, Xiang Ji ^{1,2}, Fangfang Zhang ^{1,2}, Liang Li ^{1,2} and Lan Ma ^{1,*} Embryonic Stem Cell Markers *Molecules* 2012, *17*, 6196-6236; doi:10.3390/molecules17066196

Wolfgang Huber, Anja von Heydebreck, Martin Vingron Analysis of microarray gene expression data April 2, 2003.

Yajun Liu, De Cheng, Zhenzhen Li, Xing Gao and Huayan Wang 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*, 35, 3, 693-700 (2012).

Yishai Avior,¹ Juan Carlos Biancotti,^{2,3} and Nissim Benvenisty^{1,*} TeratoScore: Assessing the Differentiation Potential of Human Pluripotent Stem Cells by Quantitative Expression Analysis of Teratomas.

APPENDIX

Program:

Java Program for Matching Marker Genes with the Genes of MicroArray Data:

```
package javaapplication77;

import java.io.FileInputStream;

import java.io.FileNotFoundException;

import java.io.IOException;

import javax.swing.JFileChooser;

import org.apache.poi.xwpf.extractor.XWPFFWordExtractor;

import org.apache.poi.xwpf.usermodel.XWPFDocument;

public class GenesComparison {

    String Store;

    String Store1;

    public void GeneList1() throws IOException {        int i;

    String str1;    String str2;    int count5=0;    int count6=0;

    try {        JFileChooser Chooser = new JFileChooser();

    int returnvalue = Chooser.showOpenDialog(null);

    if(returnvalue==JFileChooser.APPROVE_OPTION){

    XWPFDocument document = new XWPFDocument (new FileInputStream(Chooser.getSelectedFile()));

    XWPFFWordExtractor extract = new XWPFFWordExtractor(document);

    Store = extract.getText();        for(i =0; i<=170;i++)    {        str1=Store.split("\n")[i];

    str2= str1;        count5++;        count6=count5;

    }    }    } catch(FileNotFoundException jk)    {

    } catch(Exception ml)    {

    }    }    public void GeneList2() throws IOException{

    int j;    int count1=0;    String str3;    String str4;
```

```

try {      JFileChooser Chooser = new JFileChooser();

int returnvalue = Chooser.showOpenDialog(null);

if(returnvalue==JFileChooser.APPROVE_OPTION){

XWPFDocument document = new XWPFDocument (new FileInputStream(Chooser.getSelectedFile()));

XWPFWordExtractor extract = new XWPFWordExtractor(document);

Store1=extract.getText ();      }  } catch (FileNotFoundException jk) {

} catch(Exception fg) {

} }

String Str1;   String Str2;   String Str3;   String Str4;   String Str5;

int i;   int j;   int Count=0;   int Count1=0;

int Count2=0;   int y=0;

String CommonGene="A";

int count=0;   int count1=0;   int k;

public void CompareGenes() { try{ new Thread() {      public void run(){

for(;;) {      for(i=0;i<=170;i++) { Str1=Store.split("\n")[i];      Str2=Str1;

for(j=0;j<=34107;j++) {      Str3=Store1.split("\n")[j];      Str4=Str3;

if(Str2.equals(Str4)) {      Count++;      System.out.println(Str2+" MATCHED "+Str4);

for(k=j+1;k<=34107;k++) {      Str5=Store1.split("\n")[k];

if(Str4.equals(Str5)) {      count++;      } else {

} } } else {

} Count1++; }      Count2++;

System.out.println(Count+" = Number Of Total matched Gene");

System.out.println(count+" = Number Of Duplicate matched Gene");

} System.out.println(Count+" = Number Of Total Matched Genes");

System.out.println(count+" = Number Of Duplicate Matched Genes"); } } }. start();

} catch (Exception hj) {

} } public static void main(String []args) throws IOException { GenesComparison m = new GenesComparison();

m.GeneList1();      m.GeneList2();      m.CompareGenes();      } }

```


Program:

Java Program for Checking Pluripotency Level in a particular cell line.

Java Program

```
/*
 * To change this license header, choose License Headers in Project Properties.
 * To change this template file, choose Tools | Templates
 * and open the template in the editor.
 */
package javaapplication77;

import java.io.FileInputStream;
import java.io.FileNotFoundException;
import java.io.IOException;
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.PreparedStatement;
import java.sql.ResultSet;
import javax.swing.JFileChooser;
import javax.swing.JOptionPane;
import org.apache.poi.xssf.usermodel.XSSFCell;
import org.apache.poi.xssf.usermodel.XSSFRow;
import org.apache.poi.xssf.usermodel.XSSFSheet;
import org.apache.poi.xssf.usermodel.XSSFWorkbook;

public class Pluritest {
    String value=null;    String value1=null;    String value2=null;    String value3=null;    String value4=null;
    String value5=null;    String value6=null;    String value7=null;    String value8=null;
    String value9=null;    String value10=null;    String value11=null;    String value12=null;
    String Svalue1=null;    String Svalue2=null;    String Svalue3=null;    String Svalue4=null;
    String Svalue5=null;    String Svalue6=null;    String Svalue7=null;    String Svalue8=null;
    String Svalue9=null;    String Svalue10=null;    String Svalue11=null;    String Svalue12=null;
    String Nvalue1=null;    String Nvalue2=null;    String Nvalue3=null;    String Nvalue4=null;
    String Nvalue5=null;    String Nvalue6=null;    String Nvalue7=null;    String Nvalue8=null;
    String Nvalue9=null;    String Nvalue10=null;    String Nvalue11=null;    String Nvalue12=null;
    String Value1=null;    String Value2=null;    String Value3=null;    String Value4=null;
    String Value5=null;    String Value6=null;    String Value7=null;    String Value8=null;
    String Value9=null;    String Value10=null;    String Value11=null;    String Value12=null;

    public void getFile()throws IOException {
        int Count=0;
        try {
            JFileChooser Chooser = new JFileChooser();
            int returnvalue = Chooser.showOpenDialog(null);
            if(returnvalue==JFileChooser.APPROVE_OPTION)
            {
                int reply= JOptionPane.showConfirmDialog(null,"Is Data Normalized or Not?", "Question Message",JOptionPane.YES_NO_OPTION);
                if(reply==JOptionPane.YES_OPTION){
                    XSSFWorkbook workbook = new XSSFWorkbook (new FileInputStream(Chooser.getSelectedFile()));
                    XSSFSheet sheet = workbook.getSheet("Sheet1");
                    XSSFRow row = sheet.getRow(0);
                    int colnum= row.getLastCellNum();
                    int rownum = sheet.getLastRowNum()+1;
                    System.out.println(String.valueOf(colnum));
                    System.out.println(String.valueOf(rownum));
                    for(int i=1; i<=rownum;i++) { XSSFRow Row = sheet.getRow(i); for(int j=1;j<2;j++) {
                        XSSFCell cell = Row.getCell(j); value= cell.toString(); if(value.equalsIgnoreCase("POU5F1")){
                            System.out.println(value); int number = cell.getRowIndex()+1;
                            XSSFRow Row1 = sheet.getRow(number-1); XSSFRow Row2 = sheet.getRow(0);
                            for(int z=2;z<3;z++) { XSSFCell cell1=Row1.getCell(z); value1 = cell1.toString();
                                XSSFCell cell2=Row2.getCell(z); Value1 = cell2.toString();
                            }
                            for(int a=3;a<4;a++) { XSSFCell cell1=Row1.getCell(a); value2 = cell1.toString();
                                XSSFCell cell2=Row2.getCell(a); Value2 = cell2.toString();
                            }
                        }
                    }
                }
            }
        }
    }
}
```

```

    }
    for(int a=4;a<5;a++) { XSSFCell cell1=Row1.getCell(a); value3 = cell1.toString();
    XSSFCell cell2=Row2.getCell(a); Value3 = cell2.toString();
    }
    for(int a=5;a<6;a++) { XSSFCell cell1=Row1.getCell(a); value4 = cell1.toString();
    XSSFCell cell2=Row2.getCell(a); Value4 = cell2.toString();
    }
    for(int a=6;a<7;a++) { XSSFCell cell1=Row1.getCell(a); value5 = cell1.toString();
    XSSFCell cell2=Row2.getCell(a); Value5 = cell2.toString();
    }
    for(int a=7;a<8;a++) { XSSFCell cell1=Row1.getCell(a); value6 = cell1.toString();
    XSSFCell cell2=Row2.getCell(a); Value6 = cell2.toString();
    }
    for(int a=8;a<9;a++) { XSSFCell cell1=Row1.getCell(a); value7 = cell1.toString();
    XSSFCell cell2=Row2.getCell(a); Value7 = cell2.toString();
    }
    for(int a=9;a<10;a++) { XSSFCell cell1=Row1.getCell(a); value8 = cell1.toString();
    XSSFCell cell2=Row2.getCell(a); Value8 = cell2.toString();
    }
    for(int a=10;a<11;a++) { XSSFCell cell1=Row1.getCell(a); value9 = cell1.toString();
    XSSFCell cell2=Row2.getCell(a); Value9 = cell2.toString();
    }
    for(int a=11;a<12;a++) { XSSFCell cell1=Row1.getCell(a); value10 = cell1.toString();
    XSSFCell cell2=Row2.getCell(a); Value10 = cell2.toString();
    }
    for(int a=12;a<13;a++) { XSSFCell cell1=Row1.getCell(a); value11 = cell1.toString();
    XSSFCell cell2=Row2.getCell(a); Value11 = cell2.toString();
    }
    for(int a=13;a<14;a++) { XSSFCell cell1=Row1.getCell(a); value12 = cell1.toString();
    XSSFCell cell2=Row2.getCell(a); Value12 = cell2.toString();

```

```

System.out.println(Value12); } System.out.println(number); }
if(value.equalsIgnoreCase("SOX2")) { System.out.println(value);
int number = cell.getRowIndex()+1; XSSFRow Row1 = sheet.getRow(number-1);

```

```

    for(int z=2;z<3;z++) { XSSFCell cell1=Row1.getCell(z); Svalue1 = cell1.toString();
    }
    for(int a=3;a<4;a++) { XSSFCell cell1=Row1.getCell(a); Svalue2 = cell1.toString();
    }
    for(int a=4;a<5;a++) { XSSFCell cell1=Row1.getCell(a); Svalue3 = cell1.toString();
    }
    for(int a=5;a<6;a++) { XSSFCell cell1=Row1.getCell(a); Svalue4 = cell1.toString();
    }
    for(int a=6;a<7;a++) { XSSFCell cell1=Row1.getCell(a); Svalue5 = cell1.toString();
    }
    for(int a=7;a<8;a++) { XSSFCell cell1=Row1.getCell(a); Svalue6 = cell1.toString();
    }
    for(int a=8;a<9;a++) { XSSFCell cell1=Row1.getCell(a); Svalue7 = cell1.toString();
    }
    for(int a=9;a<10;a++) { XSSFCell cell1=Row1.getCell(a); Svalue8 = cell1.toString();
    }
    for(int a=10;a<11;a++) { XSSFCell cell1=Row1.getCell(a); Svalue9 = cell1.toString();
    }
    for(int a=11;a<12;a++) { XSSFCell cell1=Row1.getCell(a); Svalue10 = cell1.toString();
    }
    for(int a=12;a<13;a++) { XSSFCell cell1=Row1.getCell(a); Svalue11 = cell1.toString();
    }
    for(int a=13;a<14;a++) { XSSFCell cell1=Row1.getCell(a); Svalue12 = cell1.toString();
    }

```

```

System.out.println(number); }
if(value.equalsIgnoreCase("NANOG")) { System.out.println(value);
int number = cell.getRowIndex()+1; XSSFRow Row1 = sheet.getRow(number-1);

```

```

for(int z=2;z<3;z++) { XSSFCell cell1=Row1.getCell(z); Nvalue1 = cell1.toString();
} for(int a=3;a<4;a++) { XSSFCell cell1=Row1.getCell(a); Nvalue2 = cell1.toString();
} for(int a=4;a<5;a++) { XSSFCell cell1=Row1.getCell(a); Nvalue3 = cell1.toString();

```

```

} for(int a=5;a<6;a++) { XSSFCell cell1=Row1.getCell(a); Nvalue4 = cell1.toString();
} for(int a=6;a<7;a++) { XSSFCell cell1=Row1.getCell(a); Nvalue5 = cell1.toString();
} for(int a=7;a<8;a++) { XSSFCell cell1=Row1.getCell(a); Nvalue6 = cell1.toString();
} for(int a=8;a<9;a++) { XSSFCell cell1=Row1.getCell(a); Nvalue7 = cell1.toString();
} for(int a=9;a<10;a++) { XSSFCell cell1=Row1.getCell(a); Nvalue8 = cell1.toString();
} for(int a=10;a<11;a++) { XSSFCell cell1=Row1.getCell(a); Nvalue9 = cell1.toString();
} for(int a=11;a<12;a++) { XSSFCell cell1=Row1.getCell(a); Nvalue10 = cell1.toString();
} for(int a=12;a<13;a++) { XSSFCell cell1=Row1.getCell(a); Nvalue11 = cell1.toString();
} for(int a=13;a<14;a++) { XSSFCell cell1=Row1.getCell(a); Nvalue12 = cell1.toString();
} System.out.println(number); } }
} } else if(reply==JOptionPane.NO_OPTION) {
} } } catch(FileNotFoundException jk) { } catch(Exception lk) { }
}

float NanogGSM1854259; float NanogGSM1854260; float NanogGSM1854261; float NanogGSM1854262;
float NanogGSM1854263; float NanogGSM1854264; float NanogGSM1854265; float NanogGSM1854266;
float NanogGSM1854268; float NanogGSM1854269; float NanogGSM1854270; float NanogGSM1854266;
float Pou5f1GSM1854259; float Pou5f1GSM1854260; float Pou5f1GSM1854261; float Pou5f1GSM1854262;
float Pou5f1GSM1854263; float Pou5f1GSM1854264; float Pou5f1GSM1854265; float Pou5f1GSM1854266;
float Pou5f1GSM1854267; float Pou5f1GSM1854268; float Pou5f1GSM1854269; float Pou5f1GSM1854270;
float Sox2GSM1854259; float Sox2GSM1854260; float Sox2GSM1854261; float Sox2GSM1854262;
float Sox2GSM1854263; float Sox2GSM1854264; float Sox2GSM1854265; float Sox2GSM1854266;
float Sox2GSM1854267; float Sox2GSM1854268; float Sox2GSM1854269; float Sox2GSM1854270;

public void Fetchdata()
{ Pou5f1GSM1854259 = Float.valueOf(value1);
Pou5f1GSM1854260= Float.valueOf(value2); Pou5f1GSM1854261= Float.valueOf(value3);
Pou5f1GSM1854262= Float.valueOf(value4); Pou5f1GSM1854263= Float.valueOf(value5);
Pou5f1GSM1854264= Float.valueOf(value6); Pou5f1GSM1854265= Float.valueOf(value7);
Pou5f1GSM1854266= Float.valueOf(value8); Pou5f1GSM1854267= Float.valueOf(value9);
Pou5f1GSM1854268= Float.valueOf(value10); Pou5f1GSM1854269= Float.valueOf(value11);
Pou5f1GSM1854270= Float.valueOf(value12); NanogGSM1854259 =Float.valueOf(Nvalue1);
NanogGSM1854260 =Float.valueOf(Nvalue2); NanogGSM1854261 =Float.valueOf(Nvalue3);
NanogGSM1854262 =Float.valueOf(Nvalue4); NanogGSM1854263 =Float.valueOf(Nvalue5);
NanogGSM1854264 =Float.valueOf(Nvalue6); NanogGSM1854265 =Float.valueOf(Nvalue7);
NanogGSM1854266 =Float.valueOf(Nvalue8); NanogGSM1854267 =Float.valueOf(Nvalue9);
NanogGSM1854268 =Float.valueOf(Nvalue10); NanogGSM1854269 =Float.valueOf(Nvalue11);
NanogGSM1854270 =Float.valueOf(Nvalue12); Sox2GSM1854259= Float.valueOf(Svalue1);
Sox2GSM1854260= Float.valueOf(Svalue2); Sox2GSM1854261= Float.valueOf(Svalue3);
Sox2GSM1854262= Float.valueOf(Svalue4); Sox2GSM1854263= Float.valueOf(Svalue5);
Sox2GSM1854264= Float.valueOf(Svalue6); Sox2GSM1854265= Float.valueOf(Svalue7);
Sox2GSM1854266= Float.valueOf(Svalue8); Sox2GSM1854267= Float.valueOf(Svalue9);
Sox2GSM1854268= Float.valueOf(Svalue10); Sox2GSM1854269= Float.valueOf(Svalue11);
Sox2GSM1854270= Float.valueOf(Svalue12);
} public void CompareExpression()
{
if(Value1.contains("GSM")&&Value2.contains("GSM")&&Value3.contains("GSM")&&Value4.contains("GSM")&&Value5.contains("GSM")&&Value6.contains("GSM")&&Value7.contains("GSM")&&Value8.contains("GSM")&&Value9.contains("GSM")&&Value10.contains("GSM")&&Value11.contains("GSM")){
if(NanogGSM1854259>=5.0&&NanogGSM1854259<=9.0){
if(Pou5f1GSM1854259>=5.0&&Pou5f1GSM1854259<=9.0){
if(Sox2GSM1854259>=5.0&&Sox2GSM1854259<=9.0){
System.out.println("Range: >=0.3 to <=3.0");
System.out.println("Range: GSM1854259\nNANOG= "+NanogGSM1854259);
System.out.println("POU5F1= "+Pou5f1GSM1854259);
System.out.println("SOX2= "+Sox2GSM1854259);
System.out.println("Highly Pluripotent Cell= "+"GSM1854259");
}
}
}
} else if(NanogGSM1854259>=2.0&&NanogGSM1854259<=4.9){
if(Pou5f1GSM1854259>=2.0&&Pou5f1GSM1854259<=4.9){
if(Sox2GSM1854259>=2.0&&Sox2GSM1854259<=4.9){
System.out.println("Partial Pluripotent Cell= "+"GSM1854259");
}
}
}
} else if(NanogGSM1854259>=3.0&&NanogGSM1854259<=1.9){
if(Pou5f1GSM1854259>=3.0&&Pou5f1GSM1854259<=1.9){
if(Sox2GSM1854259>=3.0&&Sox2GSM1854259<=1.9){

```

```

        System.out.println("Range: >=-3.0 to <=1.9");
        System.out.println("Range: GSM1854259\nNANOG= "+NanogGSM1854259);
        System.out.println("POU5F1= "+Pou5f1GSM1854259);
        System.out.println("SOX2= "+Sox2GSM1854259);
        System.out.println("Less Pluripotent Cell= "+GSM1854259");
    }
}
}
if(NanogGSM1854260>=5.0&&NanogGSM1854260<=9.0){
    if(Pou5f1GSM1854260>=5.0&&Pou5f1GSM1854260<=9.0){
        if(Sox2GSM1854260>=5.0&&Sox2GSM1854260<=9.0){
            System.out.println("Range: >=0.3 to <=3.0");
            System.out.println("Range: GSM1854260\nNANOG= "+NanogGSM1854260);
            System.out.println("POU5F1= "+Pou5f1GSM1854260);
            System.out.println("SOX2= "+Sox2GSM1854260);
            System.out.println("Highly Pluripotent Cell= "+GSM1854260");
        }
    }
}
else if(NanogGSM1854260>=2.0&&NanogGSM1854260<=4.9){
    if(Pou5f1GSM1854260>=2.0&&Pou5f1GSM1854260<=4.9){
        if(Sox2GSM1854260>=2.0&&Sox2GSM1854260<=4.9){
            System.out.println("Partial Pluripotent Cell= "+GSM1854260");
        }
    }
}
else if(NanogGSM1854260>=-3.0&&NanogGSM1854260<=1.9) {
    if(Pou5f1GSM1854260>=-3.0&&Pou5f1GSM1854260<=1.9) {
        if(Sox2GSM1854260>=-3.0&&Sox2GSM1854260<=1.9) {
            System.out.println("Range: >=-3.0 to <=1.9");
            System.out.println("Range: GSM1854260\nNANOG= "+NanogGSM1854260);
            System.out.println("POU5F1= "+Pou5f1GSM1854260);
            System.out.println("SOX2= "+Sox2GSM1854260);
            System.out.println("Less Pluripotent Cell= "+GSM1854260");
        }
    }
}
}
}
if(NanogGSM1854261>=5.0&&NanogGSM1854261<=9.0){
    if(Pou5f1GSM1854261>=5.0&&Pou5f1GSM1854261<=9.0){
        if(Sox2GSM1854261>=5.0&&Sox2GSM1854261<=9.0){
            System.out.println("Range: >=0.3 to <=3.0");
            System.out.println("Range: GSM1854261\nNANOG= "+NanogGSM1854261);
            System.out.println("POU5F1= "+Pou5f1GSM1854261);
            System.out.println("SOX2= "+Sox2GSM1854261);
            System.out.println("Highly Pluripotent Cell= "+GSM1854261");
        }
    }
}
else if(NanogGSM1854261>=2.0&&NanogGSM1854261<=4.9){
    if(Pou5f1GSM1854261>=2.0&&Pou5f1GSM1854261<=4.9){
        if(Sox2GSM1854261>=2.0&&Sox2GSM1854261<=4.9){
            System.out.println("Partial Pluripotent Cell= "+GSM1854261");
        }
    }
}
else if(NanogGSM1854261>=-3.0&&NanogGSM1854261<=1.9){
    if(Pou5f1GSM1854261>=-3.0&&Pou5f1GSM1854261<=1.9){
        if(Sox2GSM1854261>=-3.0&&Sox2GSM1854261<=1.9){
            System.out.println("Range: >=-3.0 to <=1.9");
            System.out.println("Range: GSM1854261\nNANOG= "+NanogGSM1854261);
            System.out.println("POU5F1= "+Pou5f1GSM1854261);
            System.out.println("SOX2= "+Sox2GSM1854261);
            System.out.println("Less Pluripotent Cell= "+GSM1854261");
        }
    }
}
}
}
if(NanogGSM1854262>=5.0&&NanogGSM1854262<=9.0){
    if(Pou5f1GSM1854262>=5.0&&Pou5f1GSM1854262<=9.0){
        if(Sox2GSM1854262>=5.0&&Sox2GSM1854262<=9.0){
            System.out.println("Range: >=0.3 to <=3.0");
            System.out.println("Range: GSM1854262\nNANOG= "+NanogGSM1854262);

```

```

        System.out.println("POU5F1= "+Pou5f1GSM1854262);
        System.out.println("SOX2= "+Sox2GSM1854262);
        System.out.println("Highly Pluripotent Cell= "+GSM1854262");
    }
}
}else if(NanogGSM1854262>=2.0&&NanogGSM1854262<=4.9){
    if(Pou5f1GSM1854262>=2.0&&Pou5f1GSM1854262<=4.9){
        if(Sox2GSM1854262>=2.0&&Sox2GSM1854262<=4.9){
            System.out.println("Partial Pluripotent Cell= "+GSM1854262");
        }
    }
}
}else if(NanogGSM1854262>=-3.0&&NanogGSM1854262<=1.9){
    if(Pou5f1GSM1854262>=-3.0&&Pou5f1GSM1854262<=1.9){
        if(Sox2GSM1854262>=-3.0&&Sox2GSM1854262<=1.9){
            System.out.println("Range: >=-3.0 to <=1.9");
            System.out.println("Range: GSM1854262\nNANOG= "+NanogGSM1854262);
            System.out.println("POU5F1= "+Pou5f1GSM1854262);
            System.out.println("SOX2= "+Sox2GSM1854262);
            System.out.println("Less Pluripotent Cell= "+GSM1854262");
        }
    }
}
}
    if(NanogGSM1854263>=5.0&&NanogGSM1854263<=9.0){
    if(Pou5f1GSM1854263>=5.0&&Pou5f1GSM1854263<=9.0){
        if(Sox2GSM1854263>=5.0&&Sox2GSM1854263<=9.0){
            System.out.println("Range: >=0.3 to <=3.0");
            System.out.println("Range: GSM1854263\nNANOG= "+NanogGSM1854263);
            System.out.println("POU5F1= "+Pou5f1GSM1854263);
            System.out.println("SOX2= "+Sox2GSM1854263);
            System.out.println("Highly Pluripotent Cell= "+GSM1854263");
        }
    }
}
}else if(NanogGSM1854263>=2.0&&NanogGSM1854263<=4.9){
    if(Pou5f1GSM1854263>=2.0&&Pou5f1GSM1854263<=4.9){
        if(Sox2GSM1854263>=2.0&&Sox2GSM1854263<=4.9){
            System.out.println("Partial Pluripotent Cell= "+GSM1854263");
        }
    }
}
}else if(NanogGSM1854263>=-3.0&&NanogGSM1854263<=1.9){
    if(Pou5f1GSM1854263>=-3.0&&Pou5f1GSM1854263<=1.9){
        if(Sox2GSM1854263>=-3.0&&Sox2GSM1854263<=1.9){
            System.out.println("Range: >=-3.0 to <=1.9");
            System.out.println("Range: GSM1854263\nNANOG= "+NanogGSM1854263);
            System.out.println("POU5F1= "+Pou5f1GSM1854263);
            System.out.println("SOX2= "+Sox2GSM1854263);
            System.out.println("Less Pluripotent Cell= "+GSM1854263");
        }
    }
}
}
    if(NanogGSM1854264>=5.0&&NanogGSM1854264<=9.0){
    if(Pou5f1GSM1854264>=5.0&&Pou5f1GSM1854264<=9.0){
        if(Sox2GSM1854264>=5.0&&Sox2GSM1854264<=9.0){
            System.out.println("Range: >=0.3 to <=3.0");
            System.out.println("Range: GSM1854264\nNANOG= "+NanogGSM1854264);
            System.out.println("POU5F1= "+Pou5f1GSM1854264);
            System.out.println("SOX2= "+Sox2GSM1854264);
            System.out.println("Highly Pluripotent Cell= "+GSM1854264");
        }
    }
}
}else if(NanogGSM1854264>=2.0&&NanogGSM1854264<=4.9){
    if(Pou5f1GSM1854264>=2.0&&Pou5f1GSM1854264<=4.9){
        if(Sox2GSM1854264>=2.0&&Sox2GSM1854264<=4.9){
            System.out.println("Partial Pluripotent Cell= "+GSM1854264");
        }
    }
}
}
}else if(NanogGSM1854264>=-3.0&&NanogGSM1854264<=1.9){

```

```

if(Pou5f1GSM1854264>=-3.0&&Pou5f1GSM1854264<=1.9){
  if(Sox2GSM1854264>=-3.0&&Sox2GSM1854264<=1.9){
    System.out.println("Range: >=-3.0 to <=1.9");
    System.out.println("Range: GSM1854264\nNANOG= "+NanogGSM1854264);
    System.out.println("POU5F1= "+Pou5f1GSM1854264);
    System.out.println("SOX2= "+Sox2GSM1854264);
    System.out.println("Less Pluripotent Cell= "+GSM1854264");
  }
}
}
if(NanogGSM1854265>=5.0&&NanogGSM1854265<=9.0){
if(Pou5f1GSM1854265>=5.0&&Pou5f1GSM1854265<=9.0){
  if(Sox2GSM1854265>=5.0&&Sox2GSM1854265<=9.0){
    System.out.println("Range: >=0.3 to <=3.0");
    System.out.println("Range: GSM1854265\nNANOG= "+NanogGSM1854265);
    System.out.println("POU5F1= "+Pou5f1GSM1854265);
    System.out.println("SOX2= "+Sox2GSM1854265);
    System.out.println("Highly Pluripotent Cell= "+GSM1854265");
  }
}
}
}else if(NanogGSM1854265>=2.0&&NanogGSM1854265<=4.9){
if(Pou5f1GSM1854265>=2.0&&Pou5f1GSM1854265<=4.9){
  if(Sox2GSM1854265>=2.0&&Sox2GSM1854265<=4.9){
    System.out.println("Partial Pluripotent Cell= "+GSM1854265");
  }
}
}
}else if(NanogGSM1854265>=-3.0&&NanogGSM1854265<=1.9){
if(Pou5f1GSM1854265>=-3.0&&Pou5f1GSM1854265<=1.9){
  if(Sox2GSM1854265>=-3.0&&Sox2GSM1854265<=1.9){
    System.out.println("Range: >=-3.0 to <=1.9");
    System.out.println("Range: GSM1854265\nNANOG= "+NanogGSM1854265);
    System.out.println("POU5F1= "+Pou5f1GSM1854265);
    System.out.println("SOX2= "+Sox2GSM1854265);
    System.out.println("Less Pluripotent Cell= "+GSM1854265");
  }
}
}
}
if(NanogGSM1854266>=5.0&&NanogGSM1854266<=9.0){
if(Pou5f1GSM1854266>=5.0&&Pou5f1GSM1854266<=9.0){
  if(Sox2GSM1854266>=5.0&&Sox2GSM1854266<=9.0){
    System.out.println("Range: >=0.3 to <=3.0");
    System.out.println("Range: GSM1854266\nNANOG= "+NanogGSM1854266);
    System.out.println("POU5F1= "+Pou5f1GSM1854266);
    System.out.println("SOX2= "+Sox2GSM1854266);
    System.out.println("Highly Pluripotent Cell= "+GSM1854266");
  }
}
}
}else if(NanogGSM1854266>=2.0&&NanogGSM1854266<=4.9){
if(Pou5f1GSM1854266>=2.0&&Pou5f1GSM1854266<=4.9){
  if(Sox2GSM1854266>=2.0&&Sox2GSM1854266<=4.9){
    System.out.println("Partial Pluripotent Cell= "+GSM1854266");
  }
}
}
}
}else if(NanogGSM1854266>=-3.0&&NanogGSM1854266<=1.9){
if(Pou5f1GSM1854266>=-3.0&&Pou5f1GSM1854266<=1.9){
  if(Sox2GSM1854266>=-3.0&&Sox2GSM1854266<=1.9){
    System.out.println("Range: >=-3.0 to <=1.9");
    System.out.println("Range: GSM1854266\nNANOG= "+NanogGSM1854266);
    System.out.println("POU5F1= "+Pou5f1GSM1854266);
    System.out.println("SOX2= "+Sox2GSM1854266);
    System.out.println("Less Pluripotent Cell= "+GSM1854266");
  }
}
}
}
}
if(NanogGSM1854267>=5.0&&NanogGSM1854267<=9.0){
if(Pou5f1GSM1854267>=5.0&&Pou5f1GSM1854267<=9.0){

```

```

        if(Sox2GSM1854267>=5.0&&Sox2GSM1854267<=9.0){
            System.out.println("Range: >=0.3 to <=3.0");
            System.out.println("Range: GSM1854267\nNANOG= "+NanogGSM1854267);
            System.out.println("POU5F1= "+Pou5f1GSM1854267);
            System.out.println("SOX2= "+Sox2GSM1854267);
            System.out.println("Highly Pluripotent Cell= "+GSM1854267");
        }
    }
}
}else if(NanogGSM1854267>=2.0&&NanogGSM1854267<=4.9){
    if(Pou5f1GSM1854267>=2.0&&Pou5f1GSM1854267<=4.9){
        if(Sox2GSM1854267>=2.0&&Sox2GSM1854267<=4.9){
            System.out.println("Partial Pluripotent Cell= "+GSM1854267");
        }
    }
}
}else if(NanogGSM1854267>=-3.0&&NanogGSM1854267<=1.9){
    if(Pou5f1GSM1854267>=-3.0&&Pou5f1GSM1854267<=1.9){
        if(Sox2GSM1854267>=-3.0&&Sox2GSM1854267<=1.9){
            System.out.println("Range: >=-3.0 to <=1.9");
            System.out.println("Range: GSM1854267\nNANOG= "+NanogGSM1854267);
            System.out.println("POU5F1= "+Pou5f1GSM1854267);
            System.out.println("SOX2= "+Sox2GSM1854267);
            System.out.println("Less Pluripotent Cell= "+GSM1854267");
        }
    }
}
}
    if(NanogGSM1854268>=5.0&&NanogGSM1854268<=9.0){
    if(Pou5f1GSM1854268>=5.0&&Pou5f1GSM1854268<=9.0){
        if(Sox2GSM1854268>=5.0&&Sox2GSM1854268<=9.0){
            System.out.println("Range: >=0.3 to <=3.0");
            System.out.println("Range: GSM1854268\nNANOG= "+NanogGSM1854268);
            System.out.println("POU5F1= "+Pou5f1GSM1854268);
            System.out.println("SOX2= "+Sox2GSM1854268);
            System.out.println("Highly Pluripotent Cell= "+GSM1854268");
        }
    }
}
}else if(NanogGSM1854268>=2.0&&NanogGSM1854268<=4.9){
    if(Pou5f1GSM1854268>=2.0&&Pou5f1GSM1854268<=4.9){
        if(Sox2GSM1854268>=2.0&&Sox2GSM1854268<=4.9){
            System.out.println("Partial Pluripotent Cell= "+GSM1854268");
        }
    }
}
}else if(NanogGSM1854268>=-3.0&&NanogGSM1854268<=1.9){
    if(Pou5f1GSM1854268>=-3.0&&Pou5f1GSM1854268<=1.9){
        if(Sox2GSM1854268>=-3.0&&Sox2GSM1854268<=1.9){
            System.out.println("Range: >=-3.0 to <=1.9");
            System.out.println("Range: GSM1854268\nNANOG= "+NanogGSM1854268);
            System.out.println("POU5F1= "+Pou5f1GSM1854268);
            System.out.println("SOX2= "+Sox2GSM1854268);
            System.out.println("Less Pluripotent Cell= "+GSM1854268");
        }
    }
}
}
    if(NanogGSM1854269>=5.0&&NanogGSM1854269<=9.0){
    if(Pou5f1GSM1854269>=5.0&&Pou5f1GSM1854269<=9.0){
        if(Sox2GSM1854269>=5.0&&Sox2GSM1854269<=9.0){
            System.out.println("Range: >=0.3 to <=3.0");
            System.out.println("Range: GSM1854269\nNANOG= "+NanogGSM1854269);
            System.out.println("POU5F1= "+Pou5f1GSM1854269);
            System.out.println("SOX2= "+Sox2GSM1854269);
            System.out.println("Highly Pluripotent Cell= "+GSM1854269");
        }
    }
}
}else if(NanogGSM1854269>=2.0&&NanogGSM1854269<=4.9){
    if(Pou5f1GSM1854269>=2.0&&Pou5f1GSM1854269<=4.9){
        if(Sox2GSM1854269>=2.0&&Sox2GSM1854269<=4.9){
            System.out.println("Partial Pluripotent Cell= "+GSM1854269");
        }
    }
}
}
}

```

```

    }
}
}else if(NanogGSM1854269>=-3.0&&NanogGSM1854269<=1.9){
    if(Pou5f1GSM1854269>=-3.0&&Pou5f1GSM1854269<=1.9){
        if(Sox2GSM1854269>=-3.0&&Sox2GSM1854269<=1.9){
            System.out.println("Range: >=-3.0 to <=1.9");
            System.out.println("Range: GSM1854269\nNANOG= "+NanogGSM1854269);
            System.out.println("POU5F1= "+Pou5f1GSM1854269);
            System.out.println("SOX2= "+Sox2GSM1854269);
            System.out.println("Less Pluripotent Cell= "+GSM1854269");
        }
    }
}

}

if(Value12.contains("EGSM")){
    //System.out.println("Ani");
    if(NanogGSM1854270>=2.0&&NanogGSM1854270<=3.0){
        if(Pou5f1GSM1854270>=2.0&&Pou5f1GSM1854270<=3.0){
            if(Sox2GSM1854270>=2.0&&Sox2GSM1854270<=3.0){
                System.out.println("Range: >=0.3 to <=3.0");
                System.out.println("Range: GSM1854270\nNANOG= "+NanogGSM1854270);
                System.out.println("POU5F1= "+Pou5f1GSM1854270);
                System.out.println("SOX2= "+Sox2GSM1854270);
                System.out.println("Highly Pluripotent Cell= "+GSM1854270");
            } }
        }else if(NanogGSM1854270>=1.5&&NanogGSM1854270<=1.9){
            if(Pou5f1GSM1854270>=1.5&&Pou5f1GSM1854270<=1.9){
                if(Sox2GSM1854270>=1.5&&Sox2GSM1854270<=1.9){
                    System.out.println("Partial Pluripotent Cell= "+GSM1854270");
                } }
            }else if(NanogGSM1854270>=1.2&&NanogGSM1854270<=1.49){
                if(Pou5f1GSM1854270>=1.2&&Pou5f1GSM1854270<=1.49){
                    if(Sox2GSM1854270>=1.2&&Sox2GSM1854270<=1.49){
                        System.out.println("Less Pluripotent Cell= "+GSM1854270");
                    } } } } }
}

public static void main(String[] args) throws IOException{
    Pluritest test= new Pluritest();
    test.getFile();
    test.Fetchdata();
    test.CompareExpression();
}
}

```


The Quantile Function

In general, to define the quantile which corresponds to the fraction p , use linear interpolation between the two nearest p_i .

If p lies a fraction f of the way from p_i to p_{i+1} define the p th quantile to be:

$$Q(p) = (1 - f)Q(p_i) + fQ(p_{i+1})$$

As special cases, define the median and quartiles by:

$$\begin{array}{ll} \text{Median:} & Q(.5) \\ \text{Lower Quartile:} & Q(.25) \\ \text{Upper Quartile:} & Q(.75) \end{array}$$

The function Q defined in this way is called the *Quantile Function*.

To Develop an In silico approach/method which can easily determine Pluripotency of any cell especially iPSC's by using microarray data gene expression profiling in a TEXT file format.

ORIGINALITY REPORT

11 %	10 %	4 %	3 %
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	en.wikipedia.org Internet Source	3 %
2	cellnet.hms.harvard.edu Internet Source	2 %
3	Submitted to Dublin City University Student Paper	1 %
4	crbcm.blogspot.com Internet Source	1 %
5	genesdev.cshlp.org Internet Source	<1 %
6	www.ncbi.nlm.nih.gov Internet Source	<1 %
7	www.news-medical.net Internet Source	<1 %
8	dev.biologists.org Internet Source	<1 %

9	static.pubmed.gov Internet Source	<1 %
10	dalspace.library.dal.ca Internet Source	<1 %
11	Zhao, Wenxiu, Xiang Ji, Fangfang Zhang, Liang Li, and Lan Ma. "Embryonic Stem Cell Markers", <i>Molecules</i> , 2012. Publication	<1 %
12	Submitted to National College of Ireland Student Paper	<1 %
13	Jamal, Qazi Mohammad Sajid, Anupam Dhasmana, Mohtashim Lohani, Sumbul Firdaus, Md Yousuf Ansari, Ganesh Chandra Sahoo, and Shafiul Haque. "Binding Pattern Elucidation of NNK and NNAL Cigarette Smoke Carcinogens with NER Pathway Enzymes: an Onco-Informatics Study", <i>Asian Pacific Journal of Cancer Prevention</i> , 2015. Publication	<1 %
14	genewikiplus.org Internet Source	<1 %
15	www.mdpi.com Internet Source	<1 %
16	www.citeulike.org Internet Source	<1 %

17	Nestor, Michael W, and Scott A Noggle. "Standardization of human stem cell pluripotency using bioinformatics", Stem Cell Research & Therapy, 2013. Publication	<1 %
18	Submitted to Yonsei University Student Paper	<1 %
19	futurechoices.net Internet Source	<1 %
20	Submitted to University of Westminster Student Paper	<1 %
21	www.public.iastate.edu Internet Source	<1 %
22	revistes.iec.cat Internet Source	<1 %
23	www.cancerbio.net Internet Source	<1 %
24	molecularautism.biomedcentral.com Internet Source	<1 %
25	www.science.gov Internet Source	<1 %
26	www.answers.com Internet Source	<1 %

repositorio.cepal.org

27	Internet Source	<1%
28	Singh, Vimal K., Manisha Kalsan, Neeraj Kumar, Abhishek Saini, and Ramesh Chandra. "Induced pluripotent stem cells: applications in regenerative medicine, disease modeling, and drug discovery", Frontiers in Cell and Developmental Biology, 2015. Publication	<1%
29	www.scoopweb.com Internet Source	<1%
30	Yin, Fuqiang. "Upregulation of NEK2 is associated with drug resistance in ovarian cancer", Oncology Reports, 2013. Publication	<1%
31	Fatima, Azra. "Functional and molecular analysis of cardiomyocytes derived from reprogrammed pluripotent cells and embryonic stem cells", Kölner UniversitätsPublikationsServer, 2011. Publication	<1%

EXCLUDE QUOTES OFF
EXCLUDE BIBLIOGRAPHY ON

EXCLUDE MATCHES OFF