Meta-Pathway Based Analysis Of Stemness

by
Saurav Kumar Choudhary

Under the supervision of
Dr. Vibhor Kumar

Submitted in partial fulfillment of the requirements for the degree of Master of Technology, Computational Biology

Center for Computational Biology Indraprastha Institute of Information Technology - Delhi
May, 2022

Certificate

This is to certify that the thesis titled *"Meta-Pathway Based Analysis of Stemness"* being submitted by **Saurav Kumar Choudhary** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May,2022 Dr Vibhor Kumar
Department of ComputationalBiology
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgements

Firstly, I would like to express my sincere gratitude to my guide Dr.Vibhor Kumar, Without his vision and guidance, this work would never really have seen the light of the day. His advice has not only helped make this work reach unprecedented heights, but has also moulded me into a critical thinker. Also, I would like to express my sincere gratitude to Mr. Vaibhav Mittal, for his guidance, support and insightful comments throughout the project. I would like to express my sincerest gratitude to the Center of Computational Biology at Indraprastha Institute of Information Technology Delhi for giving me the opportunity to implement my thesis. It has been a wonderful experience throughout the duration.

Abstract

Recent development in the field of stem cell research field has evoked the great expectations. Researchers these days are full fledgedly working on developing methods to use self renewal potentials of stem cells in treating incurable diseases which can be a turning point in the field of modern medicine.Analysis of cells at cellular level (scRNA-seq) has given us power to explore activities happening inside the cell through gene expression. Here, we show how a slightly transformed data can provide better interpretability of results at low computational cost. These transformed data can be used to power ML algorithms with a small number of features. Pathways are better at explaining the process occurring inside a cell and therefore We aim to find a set of pathways as signature factors which are conserved across different stem cells. These can not only explain the functioning of a cell but also predict the state of a cell even with batch effects. We also showed how prediction scores of a ML model are used to derive the biological insights from the datasets.

Contents

1 Introduction **6**

1.1 Stem cell Classification . . . 7

1.2 Types of Stem Cells . . . 9

1.3 Difference Between Stem Cells and Non stem cells . . . 10

1.4 Stemness . . . 11

1.5 Methods of Detection of Stem cells . . . 12

1.6 Importance of Stem Cells . . . 13

2 Pathways Contributing to Stemness **15**

2.1 Data . . . 15

2.2 Manual Annotation . . . 16

2.3 UniPath . . . 17

2.4 Methods . . . 18

2.5 Formulation of Meta-Pathway . . . 19

2.6 LIME . . . 21

2.7 Rank Aggregation . . . 22

2.8 Results . . . 25

3 Comparison and Applications **31**

3.1 Comparison with Known Biomarkers . . . 31

3.2 Application and Biological insights . . . 33

4 Conclusion **39**

List of Figures

1.1 Classification of Stem Cells 8
1.2 Difference between stem cells and non-stem cells 11

2.1 Transformation by UniPath 17
2.2 Workflow for Formulation of Meta-Pathway 19
2.3 Contribution of top pathways for one stem cell 21
2.4 Objective function for Rank Aggregation 23
2.5 Visual Representation of Rank Aggregation results 24
2.6 Summary of Accuracy scores of all the model 26
2.7 Comparison of Random Forest model and Rank Aggregation method . 27

3.1 Comparison of stemness signatures with known gene sets on Human datasets 32
3.2 Comparison of stemness signatures with known gene sets on Mouse datasets 33
3.3 Stemness scores of liver of MCA dataset 34
3.4 Differentially expressed gene between erythroblast showing high stemness Vs normal erythroblast 35
3.5 Stemness scores of organ-lung of MCA dataset 36
3.6 Differentially expressed genes between high stemness showing lung cells Vs normal lung cells. 37
3.7 Stemness score of normal and covid infected AT2 cells 38

Chapter 1

Introduction

Human body is composed of 37.2 trillion different types of cells. They vary in shapes, size, color, function, end location, species etc but all have one common origin. This raw material of an organism's body is called stem cells. Broadly speaking these are the cells from which all other different cell types are generated. The stem cells divide and form what are called daughter cells[1]. These daughter cells can either become other stem cells or specialized cells like muscle cells, liver cells, brain cells etc.

A balance and homeostasis between the proliferation, quiescence, regeneration, repair and replacement via interaction of different genes, proteins and pathways with the micro environment is maintained by the stem cells. Only stem cells have the innate property to generate new cell types.

Properties of all stem cells:-

Key characteristics for a cell to be called a 'stem cell' are self-renewing and to become a specialized cell type as described below:

- **Self-renew:-**

  Most important and essential property of a stem cell is its potential for infinite self-renewal. This self-renewal feature means to produce its progeny exactly the same as its original cell. One can say this is a similar property as cancer cells but the cancer cell division is uncontrolled whereas the stem cell division occurs in a controlled and regulated manner.

- **Recreate functional tissues:-**

  Adult stem cells unlike pluripotent stem cells(undifferentiated) are differentiated, which means they have the ability to give rise to the specialized cell type of any tissue or organ where they reside. Based on their specific location/tissue/organ, they may adopt specific morphological features like long brain cells, flexible muscle cells etc. To be a specialized cell type, they also reflect the relevant pattern of gene expression (sometimes also called as marker genes) of that particular tissue.

Using these key properties of the stem cells stated above, common areas of active research that is going on these days involve:

- **Understand occurrence of diseases:**

  Researchers are studying the underlying occurrences of a disease by observing stem cells maturation into different cells in bones, brains, heart, muscles, skin etc in different tissues and organs. For example, studies are very active in manipulating liver stem cells to become specialized liver stem cells to treat diseases.

- **Regenerative medicine:**

  Regenerative medicine simply means generating healthy cells to replenish and replace diseased and damaged cells. Stem cells because of their self-renewal property can be guided towards modifying into specific cells. These cells in turn can be used to repair damaged tissue in a human body. Tissue and organ transplantation are common applications.

- **Testing new drugs:**

  Prior to using the new potential drug directly on the human body, scientists use appropriate types of stem cells in labs to test their drugs for safety and quality concerns. For these tests of new drugs, the cells must be reprogrammed to obtain properties of the target cells. For example, the clinical trials of diabetes drugs, creating pancreatic beta cells that could promote insulin production in labs could be a good use of stem cells.

1.1 Stem cell Classification

Potency is the ability of any stem cell to divide and differentiate into any type of specialized cell. Based on the stem cells ability to differentiate they are categorized into five different classes like totipotent, pluripotent, multipotent, oligopotent and unipotent[2].These categories are explained in more detail below with figures showing their fate.

- **Totipotent:**

  Cells that have the capability to divide and give rise to not only many differentiated cells in an organism but also to extra-embryonic tissues are called totipotent cells. They have the ability to develop and form the three primary germ layers of an early embryo. For example, when the sperm and oocyte fertilize and give rise to a single-celled embryo. It has the power to mature in an entire embryo along with all the cells it requires for its development in addition to forming a support structure to the placenta required for the fetal development[3].

- **Pluripotent:**

  Cells that possess the ability to differentiate into any one of the embryonic germ layers (ectoderm, mesoderm or endoderm) are called pluripotent stem cells. These cells divide and can turn into any type of adult stem cells to form a tissue. These cells now cannot develop into a fetal as they now lack the ability to organize into an embryo. However, in an embryo few of the inner cell mass(ICM) are pluripotent. That means they can develop into any type of somatic/germ cell type. Recent studies reveals a novel type of pluripotent stem cells called induced pluripotent stem cells(iPSCs)[4]. These cells can be created by reprogramming the somatic cells.

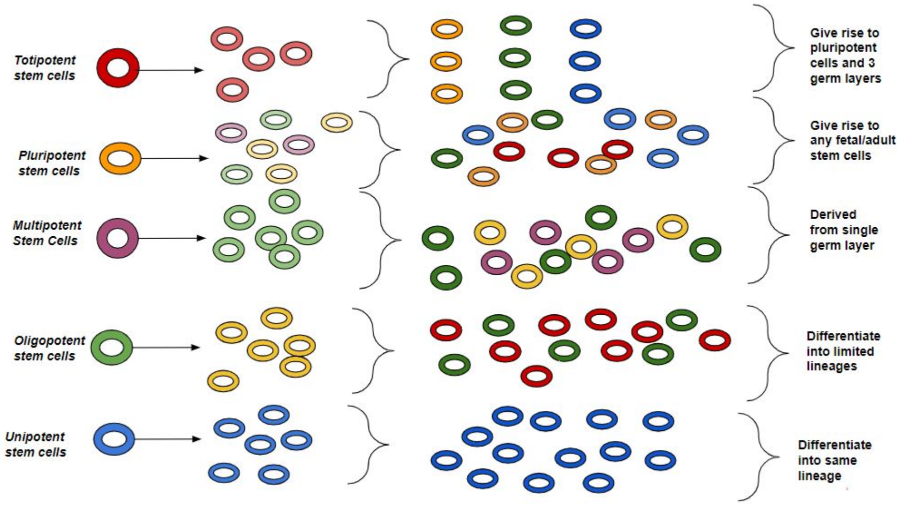

Totipotent stem cells
Give rise to pluripotent cells and 3 germ layers
Pluripotent stem cells
Give rise to any fetal/adult stem cells
Multipotent Stem Cells
Derived from single germ layer
Oligopotent stem cells
Differentiate into limited lineages
Unipotent stem cells
Differentiate into same lineage

Figure 1.1: Classification of Stem Cells

- **Multipotent:**

Cells that differentiate from any one of the germ layers are called multipotent stem cells. These can be found in most of the tissues in a body. They have the ability to differentiate into a limited number of cell types. Most common multipotent stem cells are the mesenchymal stem cells(MSCs). These originate from numerous tissues such as bone marrow, adipose tissue etc[5].These stem cells can generate various types of cell forms and can produce mesoderm-derived tissues like bone, cartilage, muscles and relevant tissues.

- **Oligopotent:**

  Stem cells with the ability that can differentiate to develop 2 or more lineages within any particular tissue and can self-regenerate are called oligopotent stem cells. For example, hematopoietic stem cells can form and differentiate into both lymphoid and myeloid lineages. Thus, these are a good example of oligopotent stem cells[6].

- **Unipotent:**

  The stem cells that have a potential of high self-renewal and narrow differentiation are the unipotent stem cells. These have the ability to give rise to cells of their own type. They are important for replacing diseased or aged cells. For example, muscle stem cells that give rise to only mature muscle cells and no other cells[7].

1.2 Types of Stem Cells

Apart from classifying stem cells based on their potency, they can also be categorized into distinct types based on their location in the body or their formation at different points of time in the body. These are comprised of embryonic stem cells that appears during the early development stages only and all other adult stem cells are tissue or organ specific based on their appearance and fetal development and remain in the body throughout the life. These types are explained in more detailed below:

- **Embryonic stem cells(ESCs) :-**

  When the embryos are 3-5 days old (called blastocyst), they are composed of cells called embryonic stem cells. At this stage, they are also called pluripotent stem cells. They have a characteristic feature to either divide into more stem cells or become any specific type of cell in the body. Undergoing through the process of embryogenesis, they can develop into one of the three germ layers-endoderm, mesoderm and ectoderm[8]. After these ESCs divide into any particular germ layer, they become multipotent stem cells. These cells are used to regenerate or repair damaged and diseased tissues/organs.

- **Adult stem cells ;-**

  Compared to embryonic stem cells, adult stem cells are limited in their ability to give rise to various cells of the body. These cells have specific specialized biological functions to perform in the organism's body.Their main feature is to maintain the tissues or organs by replacing the apoptotic cells or regenerate the damaged/injured/diseases cells of a tissue. They are very commonly used for research studies and involved in cell therapies[9]. There are many research studies going on adult stem cells as they show promising results in regenerative medicine with the potential to restore the damaged tissues when tissues transplanted in vivo.

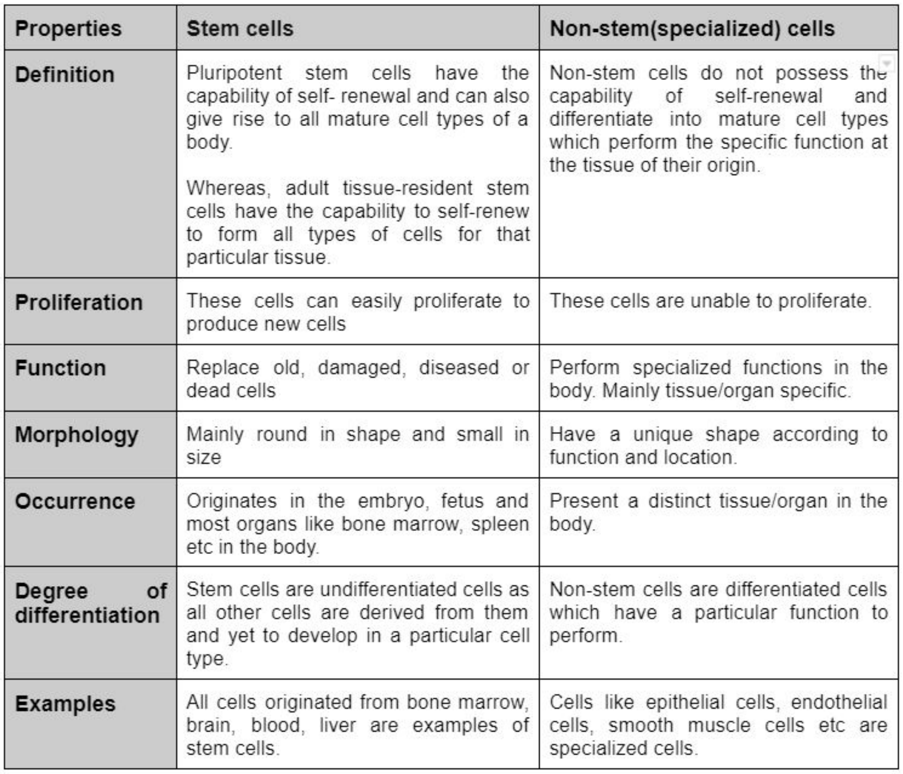
- **Perinatal stem cells :-**

  Recent discovery reveals that stem cells in amniotic fluid and umbilical cord blood have stem cells. These stem cells have the ability to change into specialized cells.

- **Induced pluripotent stem cells :-** From recent studies, researchers have successfully attempted to turn adult stem cells into pluripotent stem cells. These new cell types are the induced pluripotent stem cells(iPSCs). They have the ability to differentiate into all cells of a specialized cell in the body. In order to create these iPSCs, researchers reprogram the adult stem cells to bring back their embryonic stem-like properties.

1.3 Difference Between Stem Cells and Non stem cells

Every cell in a body does not possess the properties of stem cells. In fact, to perform special functions of a tissue, the cell needs to possess characteristic features of that tissue for the organ to work properly. Each of these categories has its own importance in maintaining homeostasis and a healthy body. Although stem cells have the ability to self renew but still lack potential to be a specialized cell. Keeping in mind all the key properties of both these categories, the below table details them:



Properties	Stem cells	Non-stem(specialized) cells
Definition	Pluripotent stem cells have the capability of self- renewal and can also give rise to all mature cell types of a body. Whereas, adult tissue-resident stem cells have the capability to self-renew to form all types of cells for that particular tissue.	Non-stem cells do not possess the capability of self-renewal and differentiate into mature cell types which perform the specific function at the tissue of their origin.
Proliferation	These cells can easily proliferate to produce new cells	These cells are unable to proliferate.
Function	Replace old, damaged, diseased or dead cells	Perform specialized functions in the body. Mainly tissue/organ specific.
Morphology	Mainly round in shape and small in size	Have a unique shape according to function and location.
Occurrence	Originates in the embryo, fetus and most organs like bone marrow, spleen etc in the body.	Present a distinct tissue/organ in the body.
Degree of differentiation	Stem cells are undifferentiated cells as all other cells are derived from them and yet to develop in a particular cell type.	Non-stem cells are differentiated cells which have a particular function to perform.
Examples	All cells originated from bone marrow, brain, blood, liver are examples of stem cells.	Cells like epithelial cells, endothelial cells, smooth muscle cells etc are specialized cells.

Figure 1.2: Difference between stem cells and non-stem cells

1.4 Stemness

Molecular functioning of what makes a stem cell perform self-renewal and generating differentiation progeny is called stemness. Stem cells in different biological environments will have different requirements to perform its function and will have different demands. These stem cells maintain a perfect balance between proliferation, quiescence and differentiation by interacting in a certain regulated manner with the microenvironment[10]. Although having such different molecular programs there are some genetic expressions that are shared by these stem cells. The key feature is that self renewal is governed by a specific set of genes that maintain certain dynamics between self-renewal and differentiation of stem cells. Under optimal circumstances, controlled changes in this balance in the signaling of certain pathways induces differentiation. But abnormal signaling cascades can initiate and develop cellular transformation and uncontrolled division. After years of research we now have started gaining in general terms that molecular biology revolves around and is closely associated with stem cells. In the near future, we might be able to see stem cells as a whole molecular identity but until then the concept of stemness simply is utility with amazing potential.

Stemness score

In this work, we have built ML models which can differentiate between the stem cells and non stem cells. Basically ML models return the probabilities of a sample belonging to a specific class which can be interpreted as stemness score here. These scores are continuous values which denote how much "stem-cell like" a cell is.

1.5 Methods of Detection of Stem cells

As we all know the function of any cell is governed by the expression of genes. Similarly, for stem cells their molecular mechanism and differentiation is also regulated by the genes. The pluripotency of various types of ESCs are maintained by a few important marker genes and few important transcription factors. For example, regulated combinations of OCT4, SOX2, KLF4 and c-MYC(OSKM) have the ability to reprogram the somatic cells and affect the pluripotency of any cell population. Each of these factors have a key role in this regulation of differentiation.

The most important out of these four is OCT4. In case there is a loss of OCT4 in embryos which lack Smad2 gene can result in premature differentiation of epiblast. This lack of Smad2 gene supports the role of OCT4 as a pluripotency marker and in ESCs maintenance[11]. SOX2 has a key function in maintaining the stemness in neural progenitor cells and in determining the fate of the cells. Loss of SOX2 can lead to NPCs differentiation[12]. Zhang et al [13] in 2010 showed that KLF4 helps OCT4 and SOX2 in regulating NANOG(gene that promotes cell induction) and recent studies say this gene could be a cancer stem cell marker in tumor progression[14]. After studying the active chromatin environment of c-MYC factor, it demonstrated its role in enhancement of cell proliferation and importance in the process of transcription(transition from initiation to elongation)[15].

Some other molecular markers are also seen to regulate stemness of a cell. SSEA1(Stage specific embryonic antigen-1) and alkaline phosphate also induce reprogramming of early stages of cells. By interacting and activating OCT4,SOX2 and NANOG, few SSEA1+ cells can facilitate in reprogramming of cells. TRA-1-60 and TRA-1-81 antigens found on the surface of human pluripotent stem cells are used commonly as markers for identifying ESCs and also expressed in teratocarcinoma cells. In 2008 Chen et al [16],showed

2 transcription regulators (p300 and Suz12) also depict the activation of some parts of protein-coding, miRNA and non-coding RNA genes in ESCs.

Stem cell culture:

When cells are grown in a laboratory, it is called ‘cell culturing’. Under suitable conditions and broth with appropriate nutrients called culture medium, stem cells can proliferate in the lab environment. Most of the stem cells in the culture petri dish attach, divide, spread and become crowded. In order to avoid overcrowding, they need to be re-plated in what is called sub-culturing. This process is carried for months and is called passage. Millions of these stem cells can be yielded from original cells. These batches are then frozen, and delivered to other cell culturing labs for further experimentation.

Next is talking of how the stem cells remain pluripotent and what triggers them to become adult stem cells. Till the time the pluripotent stem cells are in the culture dish under suitable condition and environment, they remain in the original undifferentiated form. For experiment specific, if the researchers want to generate culture for specific differentiated cell types, they change the basic chemical composition of the medium, modify the culture dish surface or at times can also alter cells by forcing expression of specific genes.

While culturing the stem cells oftentimes researchers have to do a quality check to make sure the cells are pure with no contamination. Testing the cells to make sure if they show the fundamental properties of stem cells or not is also a very key step. These test may be:

- Monitoring the rate of proliferation.
- Checking expression of marker stem cell genes to observe their function.
- By observing the chromosome of the selected cells and checking their integrity of the genome.

1.6 Importance of Stem Cells

Stem cells have lately shown promising importance for a living organism in many ways. Their self-renewal property has many benefits that can be used either to study stem cells or to use them. Some of the uses of stem cells are:

- **Regenerative organs cell types-**
  The primary feature as discussed over for a stem cell is its ability to self-renew and give rise to all kinds of specialized cell types or whole organs or repair part of damaged organs.

- **Treating diseases-**
  Although this area is still not much explored and studied yet, there has been some research trying to treat diseases like diabetes, heart disease etc.

- **Screening drugs for safety-**
  Scientists are using stem cells as the medium to test and screen novel drugs. They also develop model systems to test these drugs on and study their cell behavior. It helps them to identify any birth defects and normal growth of the cells.

- **Stem cell therapies-**
  Tissue engineering is the process of generating cells and tissues for cell-based therapies. Generally the number of adult stem cells in the tissue is limited and very small. Once removed from the body, their potential to divide becomes limited as well. This makes cell therapies very difficult and there are fewer of these cells. For stem cell therapies in diseases, scientists are able to reprogram and manipulate stem cells to initiate their ability to differentiate, transplantation and engraftment. For transplantation purposes, the stem cells must be:

  – When transplanted into the recipient, they should survive and be accepted.
  – Under the dividing phase, should be able to differentiate into the correct and desired cell type.
  – Should not be rejected by the immune system of the recipient's body.
  – Properly function for the rest of the life of recipient's.

Chapter 2

Pathways Contributing to Stemness

2.1 Data

Mus musculus

For building machine learning models using mouse datasets, samples from Mouse Cell Atlas (Han et al.[17]) were downloaded from NCBI GEO (GEO ID: GSE108097). The lab developed microwell-seq high-throughput single cell RNA sequencing technique. The study reveals expression transcriptomic profiling of various organs in mouse. 90 samples from 40 different mouse organs such as spleen, thymus, pancreas etc were studied. The dataset contained around 60,000 cells showing expression across 25133 genes. The data provided was raw annotated data of digital expression matrix in .txt format.

Another mouse dataset that used Fluidigm C1 platform for scRNA seq was considered to build the machine learning models. The platform is a combination of microfluidic technology and nanoliter-scale reactions developed by Xin et al. [18]. The data for 622 mouse samples from pancreatic tissue considering islet cells were downloaded from NCBI GEO under accession number GSE77980. The data was provided in the form of a .txt format with count matrix recorded in RPKM form. All samples combined gives a raw count for 44448 genes expressed across 663 cells.

Final mouse dataset is adopted from Streets et al. [19]., study of whole transcriptome sequencing using microfluidic platform. The authors examined the expression profile in mouse embryonic cells to perform analysis on 56 single mouse ES cells(mESCs) and 6 single mouse embryonic fibroblast(MEF) transcriptomes. The count matrices were provided in RPKM values and downloaded from NCBI GEO using accession number GSE47835. The raw count matrix was a large matrix with 23464 genes exhibiting expression across 90 cells.

For all the three above datasets, the reference genome used was the latest version(mm38) for alignment. For data filtering, only the already annotated cells were considered with unannotated cells dropped. Second layer of filtering criteria was to exclude the cells which expressed less than 700 genes. Finally after filtering there were 22094 cells from the MCA dataset used for ML model building.

Homo sapiens

Apart from mouse samples, human samples were also used for building machine learning models. One of the dataset used was from Human Cell Landscape(HCL) published by Han et al. [20]. HCL database outlines various cell types of major human organs using single cell mRNA-seq technique on microwell-seq platform. The lab also reveals single-cell hierarchy for tissue that have not been studied before. The 141 human samples for different organs were downloaded from NCBI GEO using GSE134355 accession number. There were in total 4053 cells reporting 10095 genes.

Another huge dataset was from the prominent work of Li et al[21]. under GSE81861 GEO accession number. The article describes 2 sets of datasets:(i)-1,591 single cells from 11 colorectal cancer patients and (ii)- 630 single cells from 7 cell lines. These cell lines include 55 HCT116 cells, 83A549, 23 IMR90 cells, etc mentioned on GEO. These samples were already QCed and provided in FPKM format for count matrices. The number of cells contributed from this dataset are 330 cells from 55198 genes.

Combining all the samples into one large count matrix similar to how it was done for mouse samples, the next step was to filter the unwanted cells. For the above two human sample sets, only annotated cells were considered from ML model building. Next step filtering criteria was to drop out cells that expressed less than 300 genes which counts to 4036 many cells for HCL dataset.

2.2 Manual Annotation

We used scRNA seq datasets for building machine learning models. We selected datasets where we had annotations of cell types available. All annotated cells were labeled as “Stem cell” or “Non-Stem cell” as we were training a supervised ML model for making predictions. Extensive literature mining was carried out to label the cells. For the MCA dataset there were around 700 unique cell types and for HCL dataset we had around 70 unique cell types. For datasets the gene counts were in the form of RPKM/FPKM, cell types were already labeled as stem or non stem cells.

2.3 UniPath

UniPath is an R package which is used to convert the single cell transcriptomic profiles to pathway domain. Single cell data generation is a very complex process and due to which it often leads to batch effects or variations in the data. UniPath uses an approach where it transforms the transcriptomic profiles to pathway scores using a very robust statistical method with very high accuracy. UniPath is known to be very robust in handling artifacts associated with sequencing depths and variable drop out rates.

UniPath takes input in the form of n * p where n is the gene names set as row name and p is the samples set as column names. Count can be in the form of RPKM,FPKM,TPM and UMI-counts. For converting the gene count to pathway scores it assumes that non zero values in the count matrix follow the log normal distribution and can be converted to p-values. Using Brown's method these p-values of a gene are combined for a gene set which ultimately reduces the covariation effects among genes. p -values are adjusted using Monte-Carlo approach and these adjusted p-values of a pathway are called its pathway scores[22]. Output is in the form of m *p where m is the pathway name and p is the sample name and instead of gene count there is pathway score for every cell.

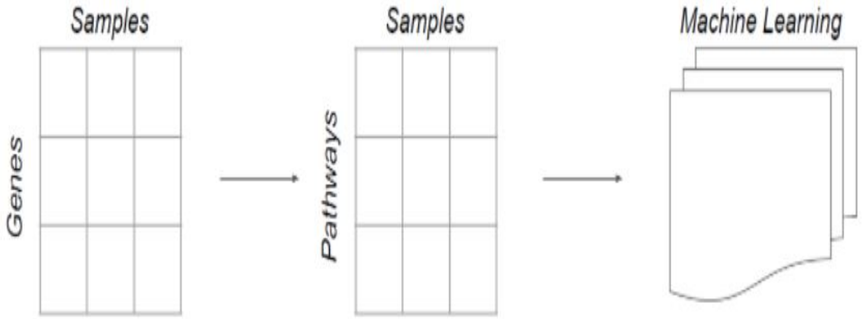


Figure 2.1: Transformation by UniPath

2.4 Methods

Genes as features - Trained on UMI count tested on UMI count

Initially machine learning models were trained using the UMI count of MCA data with genes as features. Dataset was a balanced one, consisting of 7680 stem cells and 7680 non stem cells. Hyperparameter tuning was done to train the model well, as parameters controlled the overall behavior of the models. Tuning hyperparameters takes time but it also improves the accuracy of the model. Human data having gene count in the form of UMI count was used as a test dataset, where there were 2144 cells, out of which 1077 were non stem and 1067 were stem cells. Only the common genes among MCA data and human data were taken into consideration. Best performing model was Random Forest which gave an accuracy of 55.50 %.

Genes as features - Trained on UMI count and Tested on FPKM count

Again, machine learning models were trained on the same MCA data with the same number of cell samples but this time with different numbers of genes. As for making predictions, train and test data should have the same number of features. In training data there were 15,360 cells with 8077 genes as features. Model was tested on human data with 330 cells and 8077 genes, but instead of UMI count it was FPKM values as gene count. Best performing model was a random forest with an accuracy of 50.25 %.

Pathways as features - Pathway scores calculated using UMI count

Gene expression counts were converted to pathway scores as discussed in the earlier section. UniPath conversion reduced the number of features from 25133 to 1329. Pathways related to Metabolism were removed, and we end up with 1253 pathways which were used as features for building machine learning models. Models were trained on a MCA balanced dataset having 7680 stem cells and 7680 non stem cells with 1253 pathways as features. Test dataset was human data and it was converted with the same protocol as the MCA dataset. Test data consists of 1077 non stem cells and 1067 stem cells. Random forest classification model performed best with an accuracy score of 60.35 %.

Pathways as features - Pathway scores calculated using FPKM count

Same MCA data (15320 cells and 1253 pathways) dataset was used to build the model and tested on human data which was having gene count in the form of FPKM values. There was no need to select features as these were the same in both the cases. Random forest classification model performed best with an accuracy of 50.10 %.

We will be discussing the biological significance and applications of these machine learning models in the later chapters.

2.5 Formulation of Meta-Pathway

There are many genes and pathways which are responsible for maintaining the state of a cell, but not all pathways or genes are responsible for everything that goes inside the cell. Here our aim was to find a list of top 20 pathways which are responsible for a cell to behave like a stem cell. We used ML and Rank aggregation methods to find signature factors responsible for maintaining the stemness of a cell and these can be understood better with pathways which occur inside the cell.

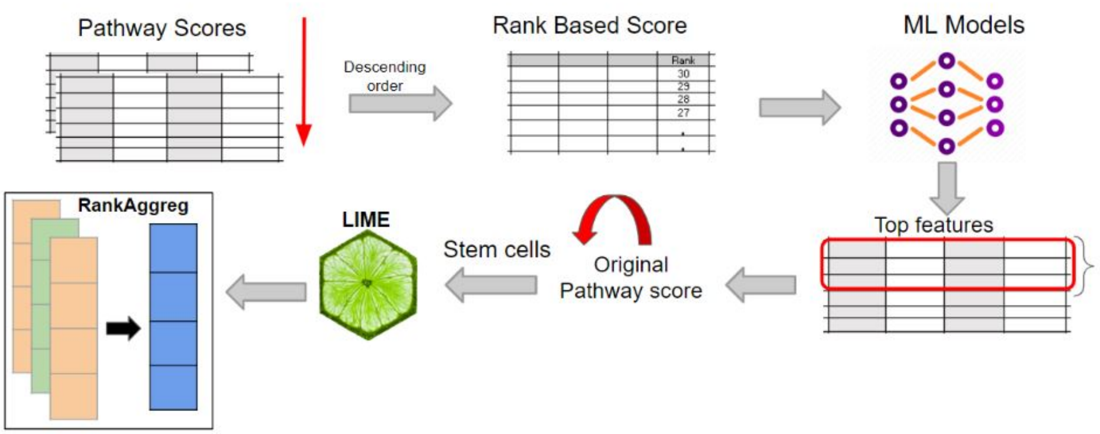


Figure 2.2: Workflow for Formulation of Meta-Pathway

Random forest is a powerful tool used with large scale datasets for classification problems. It is widely used in the bioinformatics field because of its ability to handle the “small n large p” datasets. Studies have shown that random forest is largely dependent on the type of predictor variables and its variable selection method becomes very unreliable and sometimes very misleading if there is large variation in the datasets[23].

To overcome this problem and make the variable selection unbiased, we used the quantile normalization technique. It is a widely used statistical technique, which makes the two distributions identical and comparable. This is frequently implemented in the omics field to handle batch effects[24].

We took the dataset where pathways are rownames and cells are present as columns and arranged these in a decreasing order of scores so that we can apply the quantile normalization method on it. For reducing the batch effect in the data and to make variable selection using Random forest model more reliable we replaced the pathway scores with rank based scores. For instance, for a cell top 10 pathways were given a score 30, next 10 pathways were given a score 29 and so on. This made the distribution more identical and reduced some batch effects. Directly applying a random forest model for feature selection would have given a wrong set of pathways, but rank based score made the variable very much reliable for selecting the important features.

Random Forest model was applied on these rank based scored dataset to select the top 100 pathways in an efficient manner. Replaced the rank based score with the original pathway score in the next step to make the scores biologically meaningful. In further steps, we randomly selected 200 stem cells to find out which pathways are helping these cells in maintaining the stem cell-like behavior.

For finding out the contribution of each pathway towards stemness, we used a python library called LIME which uses a model agnostic method for explaining the contribution of each feature. LIME has been discussed later in detail in this chapter. Next step was to find only those pathways which are positively regulating the stem cells, for this as_list method of the LIME library was used.

Using the LIME library we got a list of pathways for all the cells. There was a lot of variation in the sequence of pathways across the cells. Combining all these lists of pathways to make a relevant list which takes all the lists into consideration makes it a Rank Aggregation problem. We used Rank Aggreg package of R to tackle this problem which used the cross entropy method to prepare a “super list” that closely represents individual ordered lists[25]. RankAggreg package has been discussed in detail in the later portion of this chapter.

This “super-list” has 20 pathways which closely explains the stemness of a cell and represents the signature factors of stemness.

2.6 LIME

Machine learning models are widely used to make predictions about the data, but it is also known that these models are black boxes, meaning these models do not give us clear ideas about how predictions are made due to algorithmic complexity associated with it. Due to lack of interpretability it's tough to trust these models especially in the high risk fields like healthcare. Various research is going on in the field of explainable machine learning to make the output from machine learning interpretable. Various libraries and tools have been developed to overcome this problem. Some tools are specific to one type of model and some can be used with any type of model (model-agnostic)[26]. For this project purpose we used a model agnostic method called LIME.

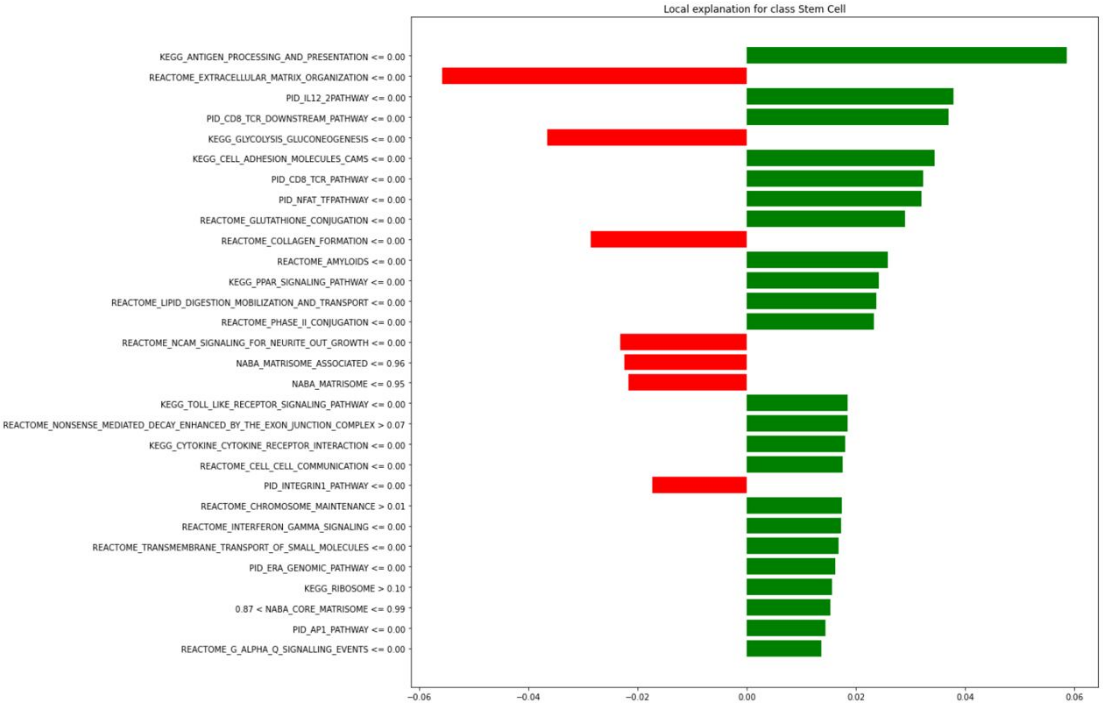


Figure 2.3: Contribution of top pathways for one stem cell

LIME stands for Local Interpretable Model Agnostic Explanations, since it is a model agnostic method, it takes any machine learning algorithm as an input and provides the detailed information about features and its contribution in making a particular prediction. It uses a surrogate model to link the features with the predictions made by black box algorithms. LIME fits a simple (interpretable) model, the explainer model, to approximate the complex model in a local region around a prediction of interest. The simple model is interpreted to identify the variables that most influence the complex model prediction.

LIME has three modules which are used with different data sets.

- lime_tabular - used for structured datasets.
- lime_text - used for text datasets.
- lime_image - used for image datasets.

Since our data was structured, we used the lime_tabular module. Lime_tabular consists of a class called LimeTabularExplainer. It takes train data, mode, labels, class names, feature names etc as an input and generates an explainer object which is used further for making interpretable predictions.

LimeTabularExplainer has a method called explain_instance() which takes a random data as an input and returns its explanation object, which contains information regarding feature contributions.

as_pyplot_figure() method can be used on an explanation object to generate a feature chart showing the contribution of each feature.

as_list() method can be applied on an explanation object to get a list of tuples which has 2 values for every data. First one represents the condition and the second value represents the contribution made by the particular feature in predicting the class of the data.

2.7 Rank Aggregation

Generally in the field of Bioinformatics and Data Science, as a final result ordered lists are generated which represent the role of genes, pathways or any other biological aspects of study. It is important to conduct the meta- analysis of these lists to find a single list which is suitable and a strong list representing all the lists. Here meta-analysis means aggregating the pathways found using LIME across the different cells. From a statistical point of view, our goal is to find a list of pathways which is the summary of distribution of observed pathways across different studies.

Rank aggregation can be treated as an optimization problem, where the goal is to find a “Super-list” which is basically a final list of multiple lists that closely represents individual ordered lists. This can be represented in a simple and intuitive form.

$$\Phi(\delta) = \sum_{i=1}^{m} w_i d(\delta, L_i),$$

Figure 2.4: Objective function for Rank Aggregation

δ represents proposed ordered list of length k = $|Li|$, wi is the important weight associated with list Li, d is a distance function and Li is the ith ordered list. The parameter ‘d’ in the above equation is the most important as it is used to measure the distance between ordered lists. The 2 most commonly used distance functions are: (a)Spearman footrule and (b)Kendall’s tau distance. Here, for the project Spearman footrule distance was used.

Spearman footrule method measures distance between ordered and ranked lists. It gives a total sum of absolute differences between ranks of elements of 2 ordered lists combined. Similarity between the lists is depicted by how smaller the value is. When comparing two top-k lists, the maximum distance could be k(k+1) which is obtained when 2 lists have no common/overlapping elements.

RankAggreg( ) function uses either CE(Cross Entropy, by default) or GA(Genetic algorithm) algorithm to perform rank aggregation.

The plot() function outputs a combination of 3 individual plots embedded in a single image.

- The first plot(top left) represents the minimum values of objective function. The y axis marks the score on each iteration shown on the x axis. The score at the 0th iteration is close to 240 which decreases at each iteration until the 8th and 9th round(slight increase). At the end of the 40th iteration, the minimum optimal score is 118.402. The score becomes stable from the 30th iteration at the score of around 118.

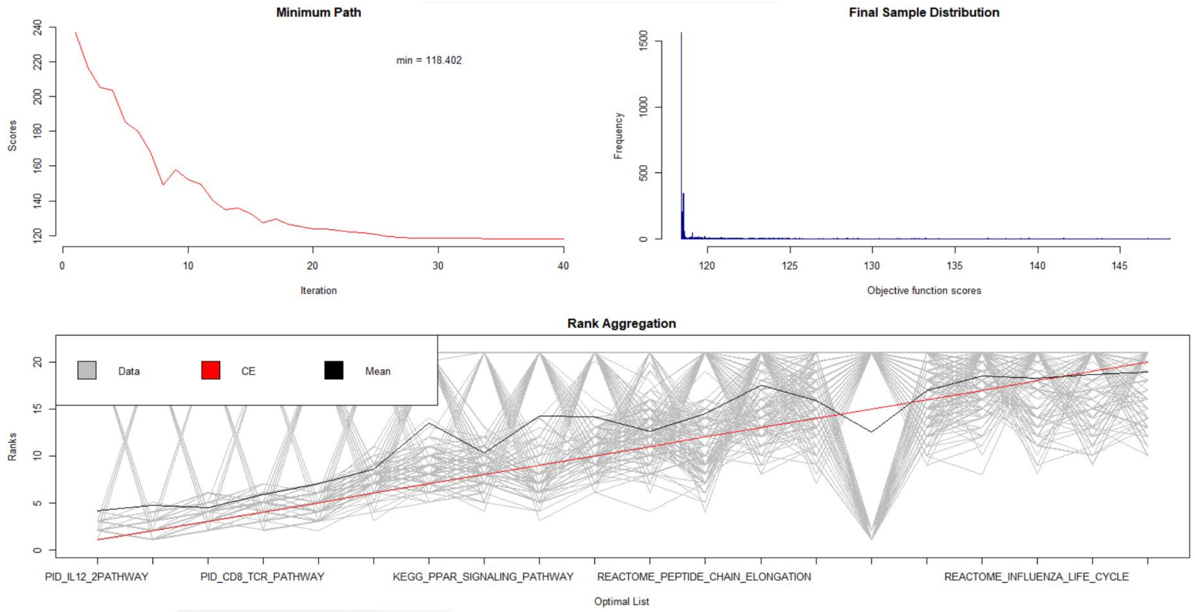


Figure 2.5: Visual Representation of Rank Aggregation results

- The second plot(top right) displays probability distribution in histogram score of objective function at the last iteration(here, 40th). The score at 118 was recorded 1500 times with very less divergence in the score frequency.
- From these 2 plots, one can get an overview of the rate of convergence(searching stopped when the minimum values of objective function does not change for a given number of iterations) and the distribution of individual lists at the end of the last iteration(40th).
- The third plot(bottom) exhibits individual ranks of lists from the data(grey lines), the final optimal rankings(red line) and mean ranking of each pathway(black line).

From the total of 200 ordered lists, we can say that in the final optimal super list, PID_IL12_2Pathway, is ranked the highest(because of its most number of occurrence in individual ordered lists and least difference in distance of ranks between individual lists) followed by PID_CD8_TCR_Pathway, KEGG_PPAR_Signaling_pathway, and so on.

2.8 Results

We investigated and compared the role of genes and pathways in making predictions across the species. For this we used four types of datasets, (1) genes as features and UMI count as gene count, (2) genes as features and FPKM/RPKM count as gene count, (3) pathways as features and pathway score calculated using UMI count and pathway as features and (4) pathway score calculated using FPKM/RPKM count. We ran the same ML workflow on all these datasets to make the inferences out of it.

We trained all the models on mouse datasets and evaluated its performance on human datasets. We found that UniPath's transformed scores are performing better than gene count in predicting the class across the species if the count matrix is UMI. Single cell transcriptomic profiles are known to have variability and artifacts associated with the data, UniPath's transformation technique not only reduces the number of features for ML algorithm but it also handles these artifacts in a robust manner. From the figure 2.6 we can see the difference of nearly 5 % in the accuracy when predicting the class using genes and pathways. This analysis shows that transforming the UMI count to pathway scores not only reduces the number of features for ML models but it also gives a slight edge in prediction and provides a better alternative which can handle the cross-platform batch effect in a better manner.

Another comparison was done between UMI count and FPKM/RPKM count to find out which can predict the class across the species with high accuracy. Our analysis shows that UMI counts are better in preserving information across species and can handle batch effects more efficiently.

ML models gave the "stemness score" for all the cells which is basically the probability score returned by the model for every cell. These scores portray the state of a cell in a more meaningful manner which can be utilized for deriving more meaningful insights from different datasets. We have discussed the application of these scores in the next chapter in a detailed manner.

We aimed to identify pathways which are responsible for proper functioning of cells and helping them to maintain the stem cell-like state. Using meta-pathway analysis we got a list of 20 pathways which we are calling stemness signatures. Pathways as signature factors help in understanding the proper functioning of a cell. Using only these pathways we were able to identify stem cells in data with very high precision even across the species as shown in the figure 2.7.

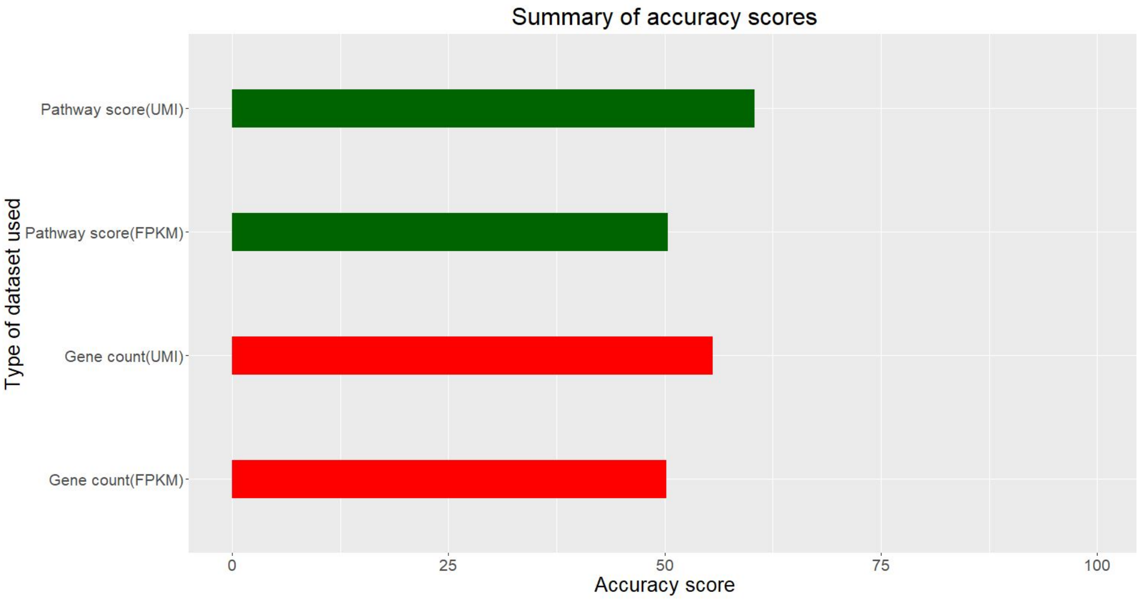


Figure 2.6: Summary of Accuracy scores of all the model

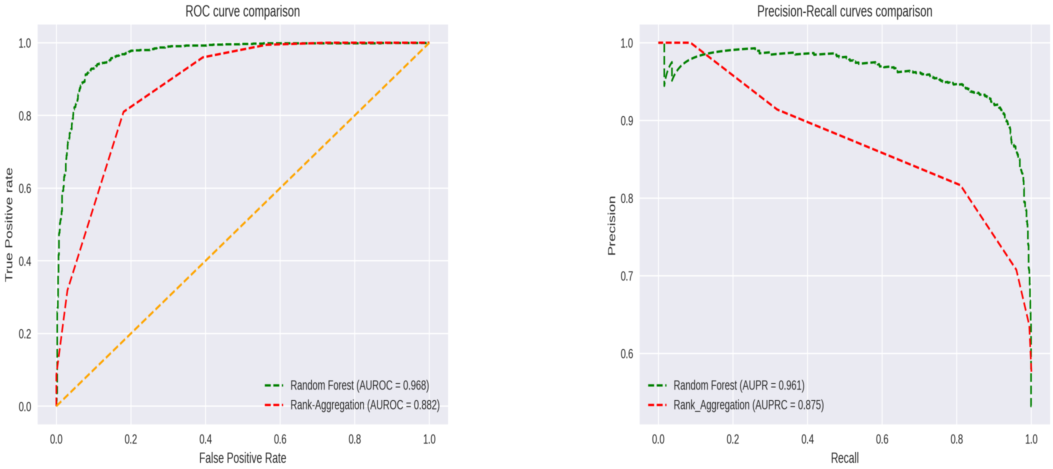
How these pathways are helping a stem cell to maintain the stemness is discussed below :- .

1. **PID_IL12_2PATHWAY**

   It includes IL12-mediated signaling events. It consists of 62 genes in the family out of which 10 genes are from cytokines and growth factors family and 16 are from cell differentiation markers family. Cytokines are known to play an important role in deriving HSCs towards division. ILs in combination with stem cell factor can lead HSCs into cell cycle. Some studies have also reported that presence of ILs slows down the differentiation rate of HSCs which ultimately preserves the self-renewal capacity of these cells[13].

2. **PID_CD8_TCR_DOWNSTREAM_PATHWAY**

   CD8 T cells play a significant role in bone marrow. CD8 T in activated form enhances the differentiation capacity of HSCs and it can also inhibit the differentiation in inactivated form which helps in maintaining the self renewal capacity



(a) auROC plot for pathway and stemness signature

(b) auPRC plot for pathway and stemness signature

Figure 2.7: Comparison of Random Forest model and Rank Aggregation method

of HSCs.[27].

3. **KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION**

   MSCs are one of the unconventional antigen presenting cells which gets stimulated after the proliferation of T cells. MSCs have power to influence both adaptive and innate immunity systems[28].

4. **PID_CD8_TCR_PATHWAY**

   CD8 cells release cytokines which help HSCs in completing the differentiation process. Genes associated with CD8 contribute significantly in the extracellular matrix pathways which maintain the stem cells placed deep inside the tissue[29].

5. **REACTOME_PHASE_II_CONJUGATION**

   Main participants of this pathway are Glucuronidation, Glutathione conjugation and Cytosolic sulfonation.Studies have shown that Glutathions acts as antioxidant and it plays crucial role in preventing stem cells from oxidative stress, improves its survival and potency.

6. **KEGG_CELL_ADHESION_MOLECULES_CAMS**

   Adhesion molecules are known to anchor the stem cell to get self renewed and helps in maintaining the potency of these cells.Integrin and cadherin sends direct signals to stem cells which help in proliferation and self-renewal[30].

7. **REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLECULES**

   ABC family proteins are major participants of this pathway and it is known that these are the proteins which are highly conserved and have very high expression in HSCs. These proteins are involved in maintaining and deciding the fate of different type of stem cells[31]

8. **KEGG_PPAR_SIGNALING_PATHWAY**

   It is very well known that PPARs belong to the nuclear receptor family and are known to regulate cell proliferation and differentiation. There are three isoforms of PPAR and out of which PPAR delta plays a very crucial role in ES differentiations[32].

9. **NABA_ECM_GLYCOPROTEINS**

   Glycans are complex carbohydrate structures which are a major component of glycoproteins. By directly analyzing the glycan profile one can find the differentiation stage of hESC. hESCs have a typical N-glycan profile and they undergo changes when there is differentiation of the cell, and this can be used as a marker for identifying hESCs[33].

10. **REACTOME_TRANSMISSION_ACROSS_CHEMICAL_SYNAPSES**

    Neurotransmitters are one of the major participants of these pathways. GABA is one of the widely studied inhibitory neurotransmitters which controls the fate of blood stem cells as they undergo differentiation. By using GABA as one of the agents there is a possibility that blood stem cells can be programmed to produce more platelets in case of patients with bleeding disorders.

11. **REACTOME_GLUTATHIONE_CONJUGATION**

    Glutathione is the non-protein thiol which functions as an antioxidant in the body. Stem cells with high levels of glutathione show higher stemness. Glutathione are required in large amounts by the stem cells to maintain its function[34].

12. **REACTOME_PEPTIDE_CHAIN_ELONGATION**

    Elongation part of translation leads to decoding of mRNA and synthesis of polypeptide chains. High translation efficiency is required by the stem cells to maintain its self renewal potency.

13. **REACTOME_NEURONAL_SYSTEM**

    Potassium channels are the major participants of the neuronal system pathway. Many researches have shown that Calcium activated potassium channels are very necessary for stem cells to proliferate[35].

14. **REACTOME_INTERFERON_GAMMA_SIGNALING**

    HSCs are regulated by IFN- in both normal as well as pathological conditions[36]. IFN- is a type of pro-inflammatory cytokine which plays a critical role in regulating the HSCs during development.

15. **REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION**

    ECM regulates the proliferation, differentiation and self renewal of stem cells. Molecules of ECM act as regulatory molecules for stem cells, based on molecular composition of these cells, ECM gets deposited accordingly to provide the best environment for stem cells to grow[37].

16. **REACTOME_IMMUNOREGULATORY_INTERACTIONS_BETWEEN_A_LYMPHOID_AND_A_NON_LYMPHOID_CELL**

    A lot of genes from clusters of differentiation families are part of this pathway and these genes get highly expressed in the field ECM related pathways which takes care of stem cells present in deep tissues.

17. **REACTOME_RESPONSE_TO_ELEVATED_PLATELET_CYTOSOLIC_CA2**

    Ca2+ plays a key role in signaling pathways at various stages of stem cell differentiation. There is much evidence which suggests that Ca2+ signaling is an important step for a cell to get differentiated and it is also required for proliferation as well and these two are important characteristics of stem cells[38].

18. **REACTOME_INFLUENZA_LIFE_CYCLE**

    Influenza induced apoptosis can activate stem cells which are in quiescent state. These apoptotic signals stimulate the proliferation of stem cells which plays an important role in tissue regeneration .

19. **REACTOME_SIGNALING_BY_GPCR**

    Many evidence suggest that signaling cascades activated by GPCR signaling are directly involved in regulating the pluripotency and differentiation of ESCs. In mouse ESCs, Gs signaling is known to promote the proliferation and pluripotency and it also has a significant impact on the pluripotent stem cells[39].

20. **KEGG_RIBOSOME**

    Highly efficient translation process is required by stem cells because there is a continuous process of self renewal and differentiation. Evidence suggests that if there is disruption in the ribosome synthesis, this may lead to hindrance in stem cell differentiation and sometimes cells may die[40].

Chapter 3

Comparison and Applications

3.1 Comparison with Known Biomarkers

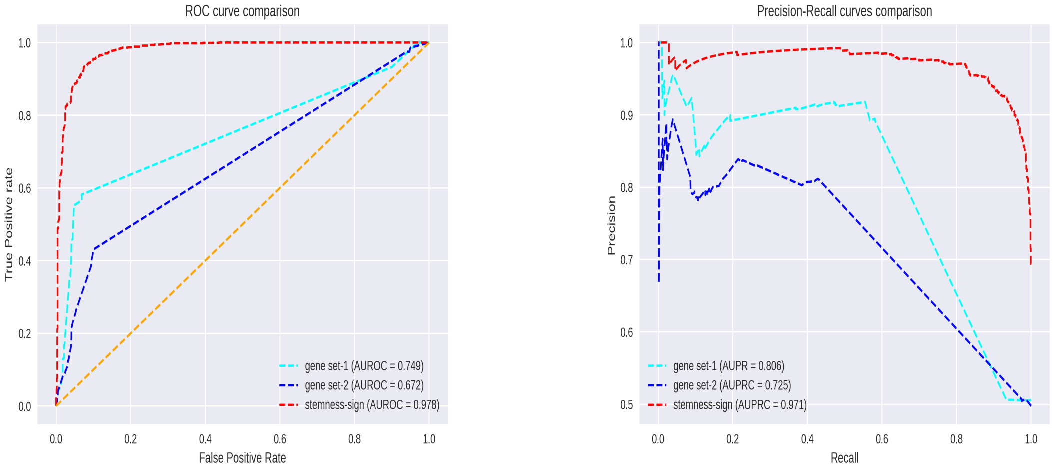
To compare the robustness of stemness signatures that we found in this work, we used the already published gene sets and markers to build ML models by using these markers as features and compared its accuracy with stemness signatures.

Benchmarking -1

Cancer stem cells are known to be subset of cancer cells. Various studies have shown their involvement in metastasis, drug resistance and relapse. Similar to normal stem cells, CSC exhibits self-renewal and differentiation[41]. Studies have shown that the degree of differentiation of a cancer can be calculated using one-class logistic regression (OCLR) which gives mRNAsi (stemness index) based on mRNA expression of stem cells. Important genes related to mRNAsi were marked by network analysis [42] have shown that mRNAsi was high in lung adenocarcinoma (LUAD) as compared to normal samples and they identified eight main gene related to mRNAsi which were - HSPA4, CDCA7, CDC20, CDK1, CLIP1, CCNB1, H2AFX and BLM. Machine learning models were built by keeping these genes and stemness signatures found in this work as features. AUROC and AUPRC curves were plotted for the comparison purpose (fig 3.1).

Benchmarking -2

Stem cells characterized by self-renewal and therapeutic resistance play crucial roles in Bladder cancer (BLCA)[43]. To find the stem cell related genes in Bladder cancer Pan et al., used the mRNAsi, weighted gene coexpression network analysis and did the



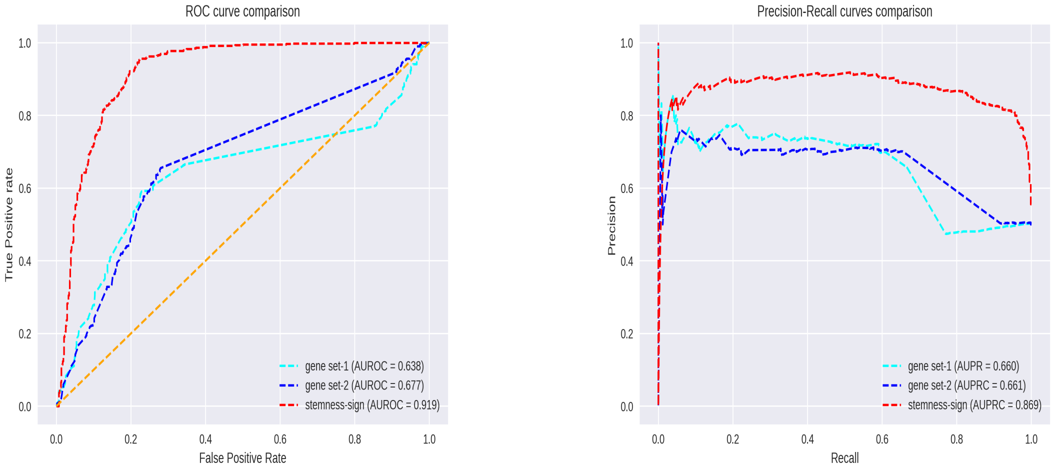
(a) auROC plot for gene sets and stemness signature

(b) auPRC plot for gene sets and stemness signature

Figure 3.1: Comparison of stemness signatures with known gene sets on Human datasets

functional annotation. They found 13 genes which were having high expression levels in the basal subtype with the highest stem cell characteristics. Major genes reported by the study were - AURKA, CDCA5, KIF11, KIF2C, KPNA2, NEK2, RACGAP1, NUSAP1, NCAPG, KIFC1, KIF18B, CDCA8 and BUB1B. ML models were built using these genes as features and compared it with models where pathways selected using ML and Rank aggregation method were features and comparison of results can be found in (fig 3.1).

These gene sets or markers were reported for humans, but since our goal included predicting the stemness across the species as well, we used these genes as features for mouse dataset as well and compared the AUROC and AUPRC (fig 3.2). Stemness signatures found in this work have outperformed the already reported gene set/markers with a very significant margin as shown in the graphs. These signature factors are able to identify stem cells with a high precision across the species as well.



(a) auROC plot for gene sets and stemness signature

(b) auPRC plot for gene sets and stemness signature

Figure 3.2: Comparison of stemness signatures with known gene sets on Mouse datasets

3.2 Application and Biological insights

As discussed earlier, Machine learning models built using genes and pathways as features were used to visualize the various datasets based on the stemness scores returned by the trained models. t-Distributed Stochastic Neighbor Embedding (tsne) plots are widely used for dimension reduction but it can also be used for visualizing large datasets. Stemness scores for MCA dataset were plotted organ-wise to check the presence of stemness in the cells of an organ.

Stemness score of Liver of Mouse cell atlas data

tSNE plots were plotted using the stemness scores returned by the model trained on pathway scores. It can be clearly seen (fig 3.3) that the liver has both types of cells, cells with low stemness score (coloured in cyan) and cells with moderate to high stemness score (yellow to red) . On manually checking it was found that apart from known stem cells there were few erythroblast cells which were having high stemness scores. Erythroblasts are nucleated cells which are commonly found in bone marrow and these cells develop into erythrocytes.

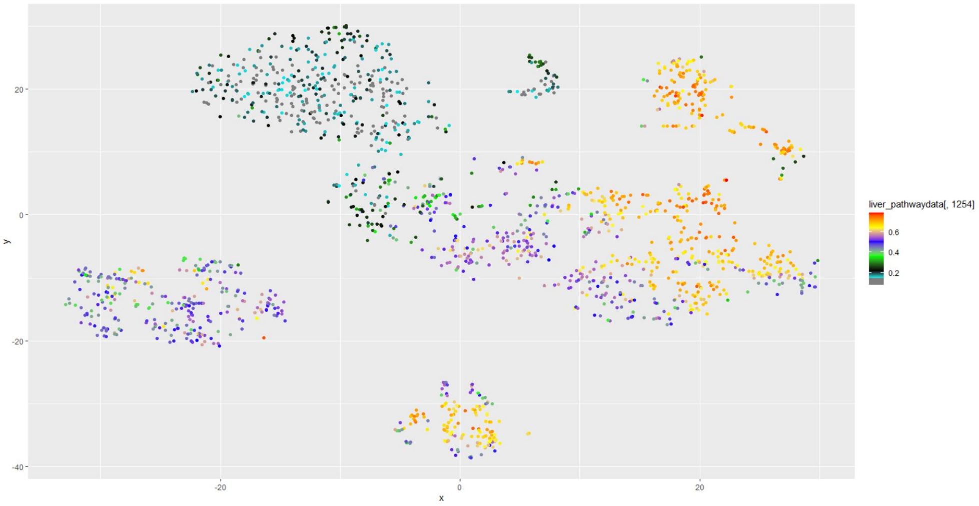


Figure 3.3: Stemness scores of liver of MCA dataset

To find out how erythroblasts with high stemness scores are different from normal erythroblasts, differential gene expression analysis was carried out between these cells(fig 3.4). The DESeq2 package of R was used to find the difference between gene expression of these two types of cells. It was found that in case of erythroblasts with high stemness, Hba-a2 is the only gene out of 25,133 variables that is most significant and has a very high log 2 fold change value. Hba-a2 is hemoglobin alpha 2, which is a protein coding gene.High level of Hba-a2 is the marker for identifying the beta-thalassemia[44]. Thalasssemia is a blood disorder where due to inefficient erythropoiesis, less number of hemoglobins are produced which leads to further complications[45].

Stemness score of organ lung of mouse cell atlas

tSNE plot for lung was also plotted which depict the stemness across the organ. We can observe there are significant amounts of yellow and red dots which denotes that there are some cells with high stemness. Out of all cells stromal cells were showing the highest stemness score. Next step was to find which genes are getting differentially expressed between cells showing high stemness and cells with low stemness. Volcano plot is showing the differentially expressed gene (fig 3.6). Out of four upregulated genes, roles of Lyz2 and Ccl6 are known in causing the stemness[46]. Lyz2 is a known late AT2 marker which is a stem cell and Ccl6 is known to promote the proliferation of Hematopoietic stem cells[47]. Role of other two upregulated genes are not known with respect to stem cells and can be explored further as a potential marker for stem cells

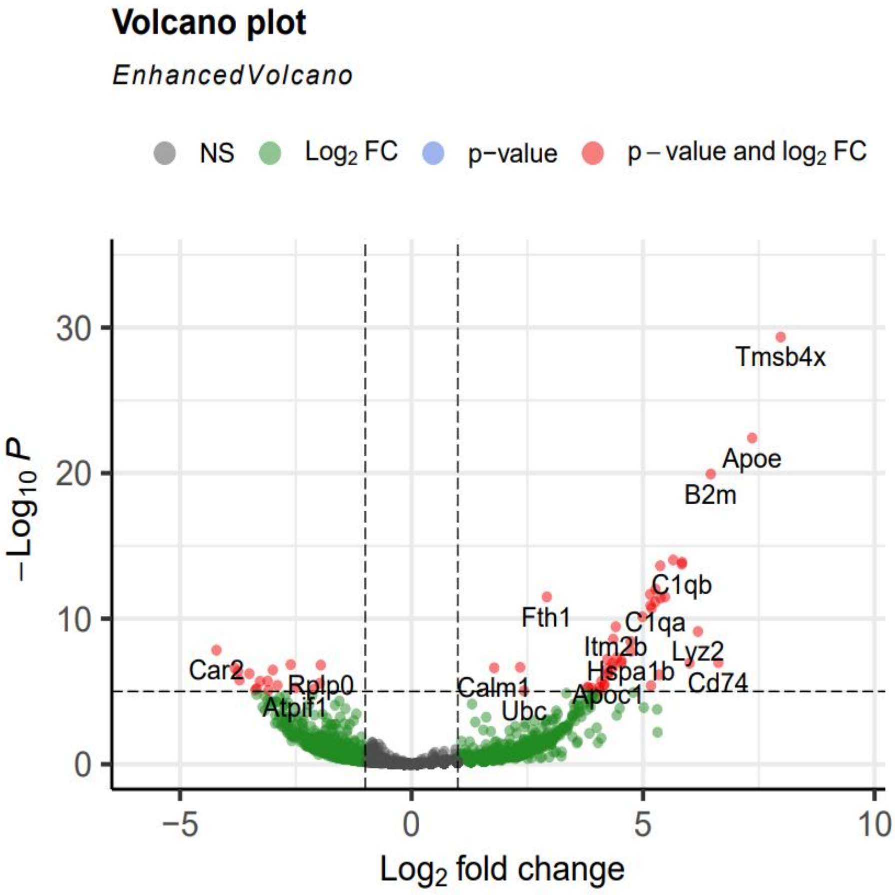


Figure 3.4: Differentially expressed gene between erythroblast showing high stemness Vs normal erythroblast

specifically for stromal cells.

Stemness score in case of covid

Some AT2 cell types of normal lung[20] were mixed with covid infected AT2 cells[48] to check the situation of AT2 cells after infection. tSNE plot was made to visualize the stemness score given by the machine learning model(fig 3.7). We can see a few red and yellow colored dots representing the normal AT2 cells. Even though data was having covid infected AT2 cells as well but model gave it low stemness score. Our finding was supported by Vakyaeva et al.[49] they showed that stem cells can be potentially

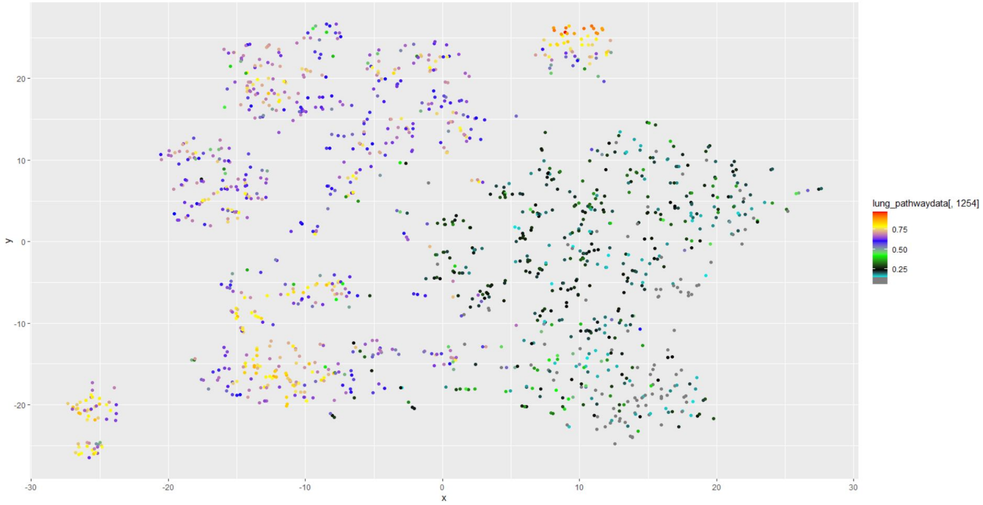


Figure 3.5: Stemness scores of organ-lung of MCA dataset

infected by SARS-CoV-2, which may lead to defects in regeneration capacity partially accounting for the severity of SARS-CoV-2 infection and its consequences.

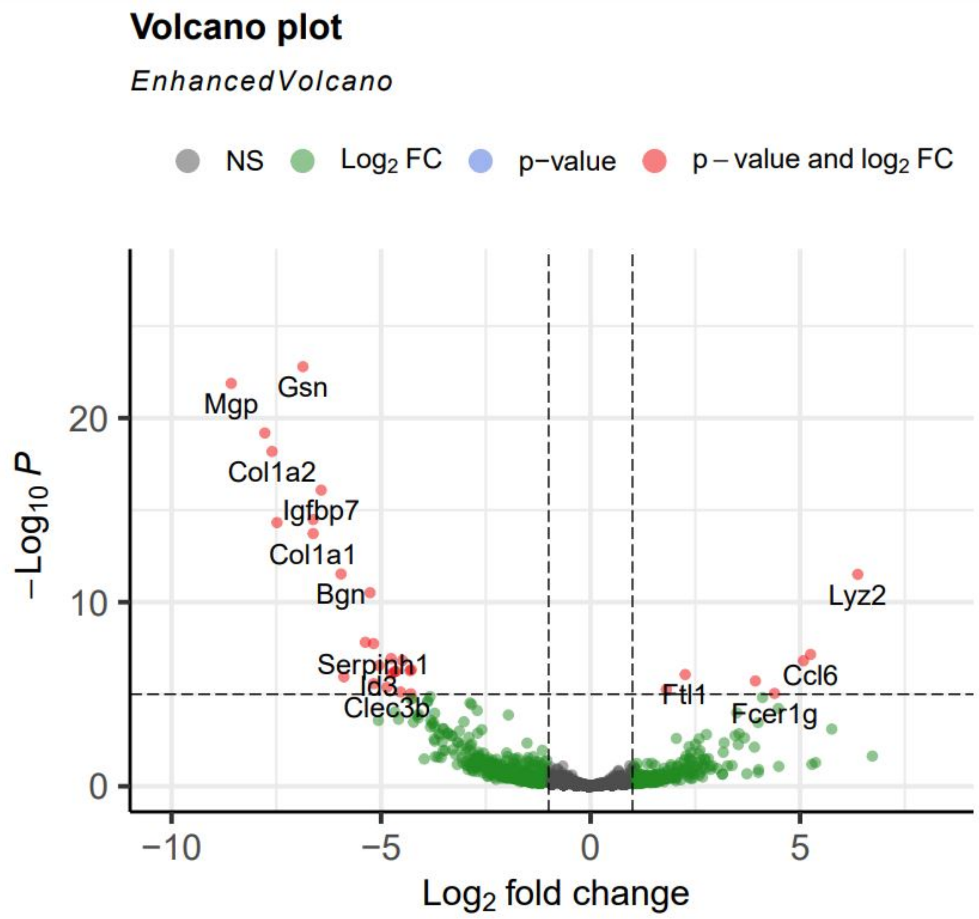


Figure 3.6: Differentially expressed genes between high stemness showing lung cells Vs normal lung cells.

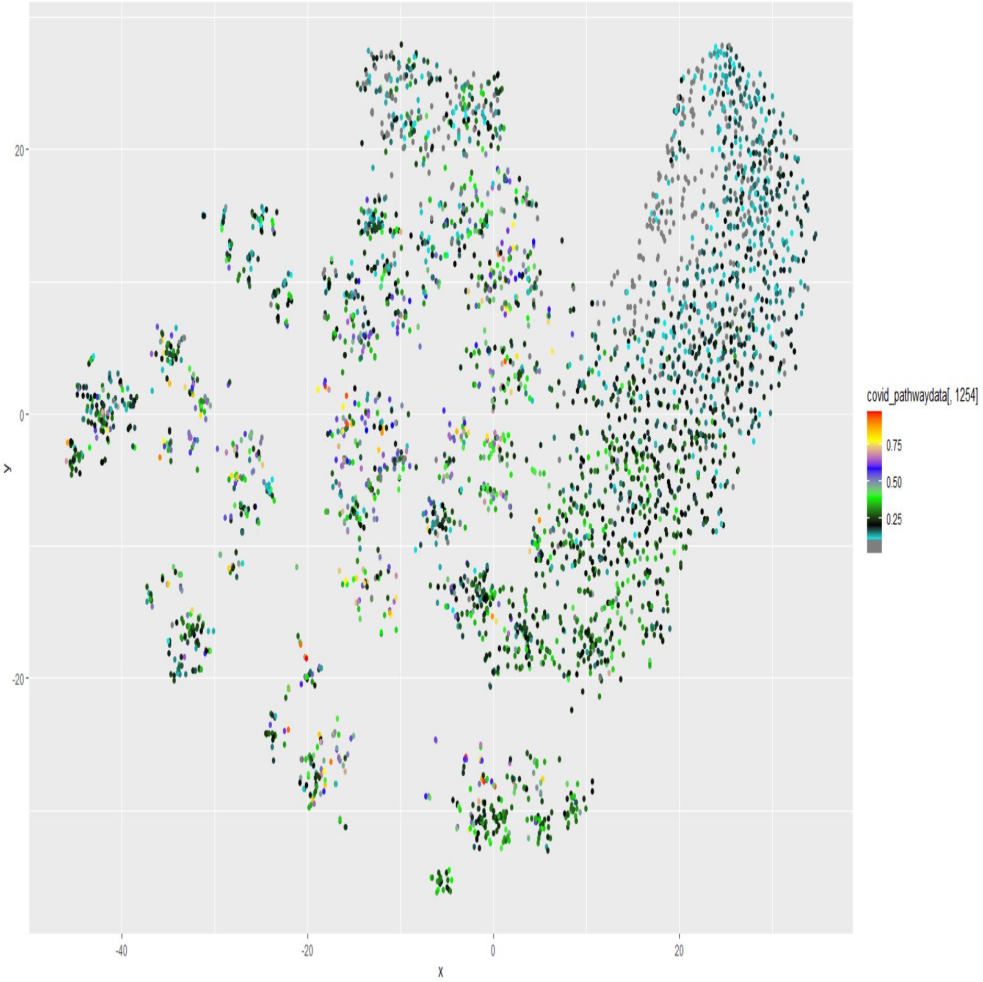


Figure 3.7: Stemness score of normal and covid infected AT2 cells

Chapter 4

Conclusion

The results of this work shows that pathway scores can be used as an alternative for gene count as features for ML models. Pathway scores can provide equivalent or sometimes better prediction as compared to gene counts and they have an added advantage of better interpretability, requiring less computational cost and robust to batch effects. Another important finding was that UMIs are better compared to FPKM or RPKM if the task is to predict across the species as UMIs are better at conserving information because they are less biased towards the gene length.

We also found a set of pathways using ML and Rank aggregation methods which are getting conserved across the batches and able to predict the class of a cell with significant accuracy. We used the pathway scores instead of gene count to find these pathways and performed quantile normalization on these scores to reduce the batch effects. In these pathways there are many pathways which are immune related like interleukin pathways, CD8 pathways,antigen processing and presentation pathways,IFN pathways and it is already known that pathways related to immune system plays a very significant role in establishing the stemness in a cell. In many pathways major participants were from cytokines growth factor family and cell differentiation marker family and genes of these families are known to play a very significant role in maintaining the stemness in a cell.

We also used the predicted value from ML models as stemness score which we used to find many important biological insights from a dataset like which organs has more stem cell, cells showing unusual stemness and may lead to some disorder in body and can also be used for finding out new marker for different types of stem cells.

This work has provided some new insights which can be implemented for different features of a cell like aging, stress and cancer too to find clusters of most significant pathways which is causing the cell to behave in a particular manner and may lead to answers to many fundamental questions in the field of translational medicine. This work also suggests that ML models can be trained and tuned to find out biological signals from datasets which have significant batch effects.

Despite showing promising results the ML models still need to be tuned and validated on different datasets to make the predictions more reliable and accurate. Also the results presented need experimental validation.

Bibliography

[1] M. Mushtaq, L. Kovalevska, S. Darekar, A. Abramsson, H. Zetterberg, V. Kashuba, G. Klein, M. Arsenian-Henriksson, and E. Kashuba, “Cell stemness is maintained upon concurrent expression of rb and the mitochondrial ribosomal protein s18-2,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, pp. 15673–15683, 2020.

[2] A. A. Shah and F. A. Khan, “Types and classification of stem cells,” in *Advances in Application of Stem Cells: From Bench to Clinics*, pp. 25–49, Springer, 2021.

[3] S. Mitalipov and D. Wolf, “Totipotency, pluripotency and nuclear reprogramming,” *Engineering of stem cells*, pp. 185–199, 2009.

[4] K. Takahashi and S. Yamanaka, “Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors,” *cell*, vol. 126, no. 4, pp. 663–676, 2006.

[5] A. Sobhani, N. Khanlarkhani, M. Baazm, F. Mohammadzadeh, A. Najafi, S. Mehdinejadiani, and F. S. Aval, “Multipotent stem cell and current application,” *Acta Medica Iranica*, pp. 6–23, 2017.

[6] C. F. B. Kim, E. L. Jackson, A. E. Woolfenden, S. Lawrence, I. Babar, S. Vogel, D. Crowley, R. T. Bronson, and T. Jacks, “Identification of bronchioalveolar stem cells in normal lung and lung cancer,” *Cell*, vol. 121, no. 6, pp. 823–835, 2005.

[7] C. F. Bentzinger, Y. X. Wang, J. von Maltzahn, and M. A. Rudnicki, “The emerging biology of muscle stem cells: Implications for cell-based therapies,” *Bioessays*, vol. 35, no. 3, pp. 231–241, 2013.

[8] S. Yao, S. Chen, J. Clark, E. Hao, G. M. Beattie, A. Hayek, and S. Ding, “Long-term self-renewal and directed differentiation of human embryonic stem cells in chemically defined conditions,” *Proceedings of the National Academy of Sciences*, vol. 103, no. 18, pp. 6907–6912, 2006.

[9] M. Chimutengwende-Gordon and W. S Khan, “Advances in the use of stem cells and tissue engineering applications in bone repair,” *Current stem cell research & therapy*, vol. 7, no. 2, pp. 122–126, 2012.

[10] D. Melton, “‘stemness’: definitions, criteria, and standards,” in *Essentials of stem cell biology*, pp. 7–17, Elsevier, 2014.

[11] W. R. Waldrip, E. K. Bikoff, P. A. Hoodless, J. L. Wrana, and E. J. Robertson, “Smad2 signaling in extraembryonic tissues determines anterior-posterior polarity of the early mouse embryo,” *Cell*, vol. 92, no. 6, pp. 797–808, 1998.

[12] V. Graham, J. Khudyakov, P. Ellis, and L. Pevny, “Sox2 functions to maintain neural progenitor identity,” *Neuron*, vol. 39, no. 5, pp. 749–765, 2003.

[13] S. Zhang, M. Morita, Z. Wang, J. Ooehara, S. Zhang, M. Xie, H. Bai, W. Yu, X. Wang, F. Dong, *et al.*, “Interleukin-12 supports in vitro self-renewal of long-term hematopoietic stem cells,” *Blood Science*, vol. 1, no. 01, pp. 92–101, 2019.

[14] P. Vasefifar, R. Motafakkerazad, L. A. Maleki, S. Najafi, F. Ghrobaninezhad, B. Najafzadeh, H. Alemohammad, M. Amini, A. Baghbanzadeh, and B. Baradaran, “Nanog, as a key cancer stem cell marker in tumor progression,” *Gene*, vol. 827, p. 146448, 2022.

[15] R. Schmidt and K. Plath, “The roles of the reprogramming factors oct4, sox2 and klf4 in resetting the somatic cell epigenome during induced pluripotent stem cell generation,” *Genome biology*, vol. 13, no. 10, pp. 1–11, 2012.

[16] Y. Chen, Y. Niu, Y. Li, Z. Ai, Y. Kang, H. Shi, Z. Xiang, Z. Yang, T. Tan, W. Si, *et al.*, “Generation of cynomolgus monkey chimeric fetuses using embryonic stem cells,” *Cell Stem Cell*, vol. 17, no. 1, pp. 116–124, 2015.

[17] X. Han, R. Wang, Y. Zhou, L. Fei, H. Sun, S. Lai, A. Saadatpour, Z. Zhou, H. Chen, F. Ye, *et al.*, “Mapping the mouse cell atlas by microwell-seq,” *Cell*, vol. 172, no. 5, pp. 1091–1107, 2018.

[18] Y. Xin, J. Kim, M. Ni, Y. Wei, H. Okamoto, J. Lee, C. Adler, K. Cavino, A. J. Murphy, G. D. Yancopoulos, *et al.*, “Use of the fluidigm c1 platform for rna sequencing of single mouse pancreatic islet cells,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 12, pp. 3293–3298, 2016.

[19] A. M. Streets, X. Zhang, C. Cao, Y. Pang, X. Wu, L. Xiong, L. Yang, Y. Fu, L. Zhao, F. Tang, *et al.*, "Microfluidic single-cell whole-transcriptome sequencing," *Proceedings of the National Academy of Sciences*, vol. 111, no. 19, pp. 7048–7053, 2014.

[20] X. Han, Z. Zhou, L. Fei, H. Sun, R. Wang, Y. Chen, H. Chen, J. Wang, H. Tang, W. Ge, *et al.*, "Construction of a human cell landscape at single-cell level," *Nature*, vol. 581, no. 7808, pp. 303–309, 2020.

[21] H. Li, E. T. Courtois, D. Sengupta, Y. Tan, K. H. Chen, J. J. L. Goh, S. L. Kong, C. Chua, L. K. Hon, W. S. Tan, *et al.*, "Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors," *Nature genetics*, vol. 49, no. 5, pp. 708–718, 2017.

[22] S. Chawla, S. Samydurai, S. L. Kong, Z. Wu, Z. Wang, W. L. Tam, D. Sengupta, and V. Kumar, "Unipath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles," *Nucleic acids research*, vol. 49, no. 3, pp. e13–e13, 2021.

[23] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, pp. 1–21, 2007.

[24] Y. Guo, S. Zhao, P.-F. Su, C.-I. Li, F. Ye, C. R. Flynn, and Y. Shyr, "Statistical strategies for micrornaseq batch effect reduction," *Translational cancer research*, vol. 3, no. 3, p. 260, 2014.

[25] V. Pihur, S. Datta, and S. Datta, "Rankaggreg, an r package for weighted rank aggregation," *BMC bioinformatics*, vol. 10, no. 1, pp. 1–10, 2009.

[26] Sunny, "Lime." https://coderzcolumn.com/tutorials/machine-learning/how-to-use-lime-to-understand-sklearn-models-predictions, oct 2021.

[27] S. Geerman, G. Brasser, S. Bhushal, F. Salerno, N. A. Kragten, M. Hoogenboezem, G. de Haan, M. C. Wolkers, M. F. Pascutti, and M. A. Nolte, "Memory cd8+ t cells support the maintenance of hematopoietic stem cells in the bone marrow," *haematologica*, vol. 103, no. 6, p. e230, 2018.

[28] K. M. Van Megen, E.-J. T. Van't Wout, J. Lages Motta, B. Dekker, T. Nikolic, and B. O. Roep, "Activated mesenchymal stromal cells process and present antigens regulating adaptive immunity," *Frontiers in immunology*, p. 694, 2019.

[29] S. Nallanthighal, J. P. Heiserman, and D.-J. Cheon, “The role of the extracellular matrix in cancer stemness,” *Frontiers in cell and developmental biology*, vol. 7, p. 86, 2019.

[30] S. Chen, M. Lewallen, and T. Xie, “Adhesion in the stem cell niche: biological roles and regulation,” *Development*, vol. 140, no. 2, pp. 255–265, 2013.

[31] M. Raaijmakers, “Atp-binding-cassette transporters in hematopoietic stem cells and their utility as therapeutical targets in acute and chronic myeloid leukemia,” *Leukemia*, vol. 21, no. 10, pp. 2094–2102, 2007.

[32] M. Y. Lee, Y. J. Lee, Y. H. Kim, S. H. Lee, J. H. Park, M. O. Kim, H. N. Suh, J. M. Ryu, S. P. Yun, M. W. Jang, *et al.*, “Role of peroxisome proliferator-activated receptor (ppar) δ in embryonic stem cell proliferation,” *International Journal of Stem Cells*, vol. 2, no. 1, p. 28, 2009.

[33] T. Satomaa, A. Heiskanen, M. Mikkola, C. Olsson, M. Blomqvist, M. Tiittanen, T. Jaatinen, O. Aitio, A. Olonen, J. Helin, *et al.*, “The n-glycome of human embryonic stem cells,” *BMC cell biology*, vol. 10, no. 1, pp. 1–18, 2009.

[34] E. M. Jeong, J.-H. Yoon, J. Lim, J.-W. Shin, A. Y. Cho, J. Heo, K. B. Lee, J.-H. Lee, W. J. Lee, H.-J. Kim, *et al.*, “Real-time monitoring of glutathione in living cells reveals that high glutathione levels are required to maintain stem cell function,” *Stem cell reports*, vol. 10, no. 2, pp. 600–614, 2018.

[35] G.-R. Li and X.-L. Deng, “Functional ion channels in stem cells,” *World journal of stem cells*, vol. 3, no. 3, p. 19, 2011.

[36] D. E. Morales-Mantilla and K. Y. King, “The role of interferon-gamma in hematopoietic stem cell development, homeostasis, and disease,” *Current stem cell reports*, vol. 4, no. 3, pp. 264–271, 2018.

[37] F. Gattazzo, A. Urciuolo, and P. Bonaldo, “Extracellular matrix: a dynamic microenvironment for stem cell niche,” *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1840, no. 8, pp. 2506–2519, 2014.

[38] F. M. Tonelli, A. K. Santos, D. A. Gomes, S. L. d. Silva, K. N. Gomes, L. O. Ladeira, and R. R. Resende, “Stem cells and calcium signaling,” *Calcium signaling*, pp. 891–916, 2012.

[39] J. R. Lynch and J. Y. Wang, "G protein-coupled receptor signaling in stem cells and cancer," *International journal of molecular sciences*, vol. 17, no. 5, p. 707, 2016.

[40] Z. Han, Q. Zhang, Y. Zhu, J. Chen, and W. Li, "Ribosomes: an exciting avenue in stem cell research," *Stem cells international*, vol. 2020, 2020.

[41] S. Singh and S. Chellappan, "Lung cancer stem cells: Molecular features and therapeutic targets," *Molecular aspects of medicine*, vol. 39, pp. 50–60, 2014.

[42] M. Zhao, Z. Chen, Y. Zheng, J. Liang, Z. Hu, Y. Bian, T. Jiang, M. Li, C. Zhan, M. Feng, *et al.*, "Identification of cancer stem cell-related biomarkers in lung adenocarcinoma by stemness index and weighted correlation network analysis," *Journal of cancer research and clinical oncology*, vol. 146, no. 6, pp. 1463–1472, 2020.

[43] S. Pan, Y. Zhan, X. Chen, B. Wu, and B. Liu, "Identification of biomarkers for controlling cancer stem cell characteristics in bladder cancer by network analysis of transcriptome data stemness indices," *Frontiers in oncology*, p. 613, 2019.

[44] S. Colaco, R. Colah, and A. Nadkarni, "Significance of borderline hba2 levels in β thalassemia carrier screening," *Scientific Reports*, vol. 12, no. 1, pp. 1–10, 2022.

[45] Bajwa, "Thalassemia." `https://www.ncbi.nlm.nih.gov/books/NBK545151/`, nov 2021.

[46] T. J. Desai, D. G. Brownfield, and M. A. Krasnow, "Alveolar progenitor and stem cells in lung development, renewal and cancer," *Nature*, vol. 507, no. 7491, pp. 190–194, 2014.

[47] F. Li, X. Du, F. Lan, N. Li, C. Zhang, C. Zhu, X. Wang, Y. He, Z. Shao, H. Chen, *et al.*, "Eosinophilic inflammation promotes ccl6-dependent metastatic tumor growth," *Science Advances*, vol. 7, no. 22, p. eabb5943, 2021.

[48] T. M. Delorey, C. G. Ziegler, G. Heimberg, R. Normand, Y. Yang, Å. Segerstolpe, D. Abbondanza, S. J. Fleming, A. Subramanian, D. T. Montoro, *et al.*, "Covid-19 tissue atlases reveal sars-cov-2 pathology and cellular targets," *Nature*, vol. 595, no. 7865, pp. 107–113, 2021.

[49] A. A. Valyaeva, A. A. Zharikova, A. S. Kasianov, Y. S. Vassetzky, and E. V. Sheval, "Expression of sars-cov-2 entry factors in lung epithelial stem cells and its potential implications for covid-19," *Scientific Reports*, vol. 10, no. 1, pp. 1–8, 2020.