Analysis of transcriptomic differences between iPSC and ESC

Andrea Ghigi, Alessandro Marincioni, Luca Fusar Bassini

ABSTRACT

Induced pluripotent stem cells (iPSCs) are a type of stem cell that can be generated directly from a somatic cell, closely resembling embryonic stem cells (ESCs). In our project, different methods were explored to classify and distinguish ESCs and iPSCs on their gene expression, to see how much the source tissue influences the resulting iPSC and which genes differentiate the two classes.

Gene expression data was collected in the form of Robust Multiarray Average (RMA) scores from microarray datasets found on the Gene Expression Omnibus (GEO) database. The obtained dataset was preprocessed to reduce the 'batch-effect', often observed between RMA measures of different microarray experiments, and to reduce feature space dimension. Source tissue classification was done by means of kNN. Linear discriminant analysis (LDA) models were developed to compare the predictive capabilities of different sets of differentially regulated genes.

Our results indicate gene expression differences between iPSCs of different origin tissue, as the classification methods we used achieve accuracies of 87.5% and 93.75%. Comparing the performance of different LDA models we found that just two genes are sufficient to separate ESCs and iPSCs with 85% accuracy and using more does not improve the performance, showing that iPSCs are imprecise replicas of ESCs.

Introduction

Embryonic stem cells (ESCs) are the sole stem cells in the human body characterized by pluripotency, namely the ability to differentiate into any cell type of the adult organism. Pluripotency makes ESCs an extremely valuable tool both for basic research and for clinical applications: in fact, treatments based on ESCs are under development to cure conditions including Parkinson's disease[1], blindness[2] and spinal cord injuries[3]. However, several technical and ethical questions have been raised on ESCs collection and utilization[4].

Successful generation of human induced pluripotent stem cells (iPSCs) deriving from somatic cells could provide promising substitutes for ESCs, which still represent the “gold standard” for regenerative medicine[5]. Transcriptomic studies have revealed similar patterns of gene expression between ESCs and iPSCs, but some fundamental differences remain that distinguish them[6,7]. Here, we collect microarray data from multiple experiments to compare the transcriptome of several different lines of ESCs and iPSCs for cluster analysis and classification.

Materials and Methods

Collection of the data and construction of the dataset

Initially, we identified a total of 230 transcriptomes in Gene Expression Omnibus (GEO)[8], an open access database of transcriptomic data, and we wrote an algorithm to extract them automatically. However, the measure units of gene expression were not standardized, therefore we retained only those samples which shared the most frequent one, that is Robust Multiarray Average (RMA)[9].
Among the 72 remaining samples, 16 were ESCs, whereas 56 were iPSCs. The original number of features was 33297, corresponding to all the DNA probes in the microarrays. The dataset contained a total of 569 missing values dispersed in 75 features, which were simply removed from the dataset. However, not all the over 33000 remaining features

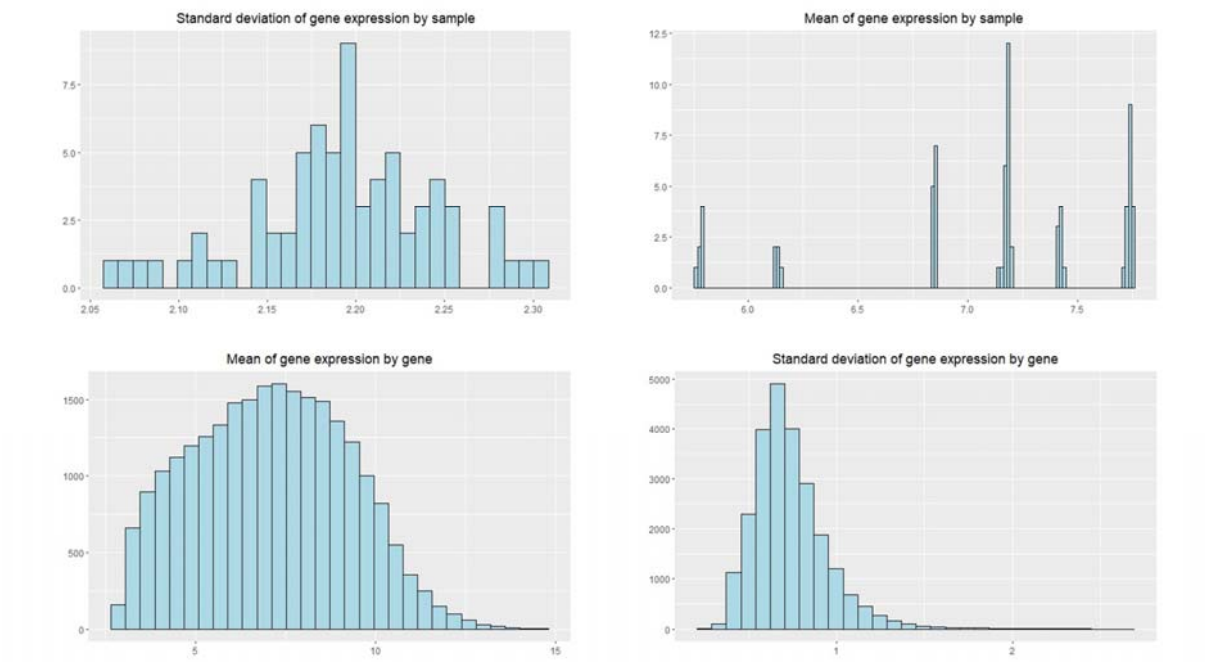


Figure 1: Histograms of means and standard deviations of gene expression both by sample and by gene

corresponded to coding genes. Therefore, we used a tool called gProfiler[10] to identify all the probes matching genes and to convert them to their respective gene names according to the standardized ENSEMBL[11] nomenclature. All the probes which were not recognized by gProfiler did not correspond to coding genes and were discarded. The final number of remaining features was 24249.

A strong experiment effect compromises the comparability of the transcriptomes

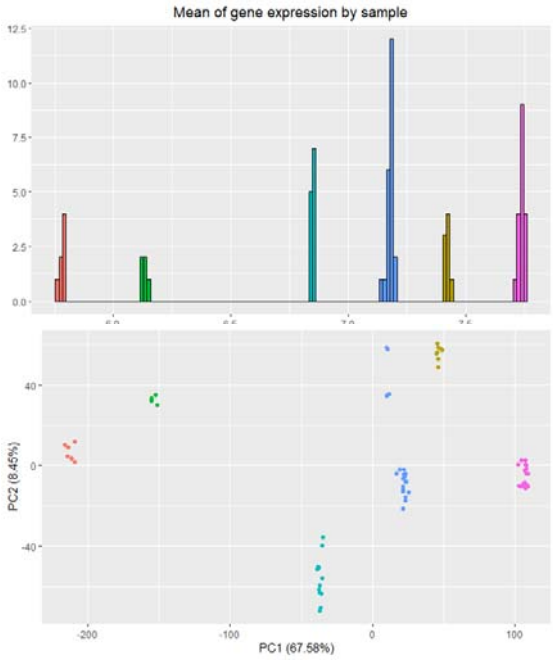


Figure 2: Histogram of the means of gene expression coloured by experiment and PCA where each colour represents a different experiment.

After completing our dataset, we decided to carry out some preliminary tests. We plotted the means and standard deviations of gene expression both by sample and by gene (fig. 1). The distribution does not strongly deviate from normality when looking at means and standard deviations of gene expression by gene, but on the contrary the histogram of the means of gene expression by sample shows 6 well-separated spikes, each of which could represent a distinct gaussian curve. Since 6 is the number of experiments from which our 72 samples were drawn, we hypothesised that each experiment determined a different mean of gene expression. This is confirmed by the histogram and the PCA coloured by experiment (fig. 2). It is evident from the PCA that most of the variability within the data is explained by the experiment. There is plenty of literature dealing with the so-called "batch-effect"[12,13], however, the level of inter-experimental variability in this case is so overwhelming that it can hardly be ascribed to a simple difference in experimental executions. The main cause of variability was likely generated by the RMA pre-processing, which entails a step of normalization that could

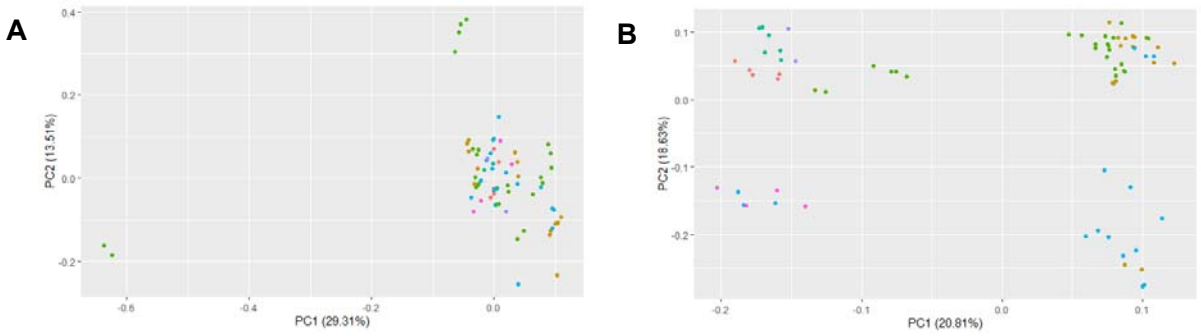


Figure 3: (A) PCA obtained after scaling with z-score normalization. (B) PCA obtained after scaling by experiment.

compromise the comparability of data from different experiments[9]. Thus, we deem it absolutely depreciable that the majority of the microarray data we found in GEO was already pre-processed. In order to eliminate the effect of the experiment, we considered two different scaling methods. The first was a simple z-score normalization applied to each sample; the second method instead, involved fitting an ANOVA to estimate the mean of gene expression in each experiment, which then was subtracted from the gene expression of their relative samples[13]. The PCA obtained after scaling by sample (fig. 3A) shows that most of the experiment-effect has been removed and that natural clusters are discernible. The PCA in figure 3B, where data are scaled by experiment, shows that samples form a cloud where it is hard to recognize natural clusters. Therefore, z-score normalization was finally applied to the dataset.

Removal of outliers and feature space dimension reduction with GO

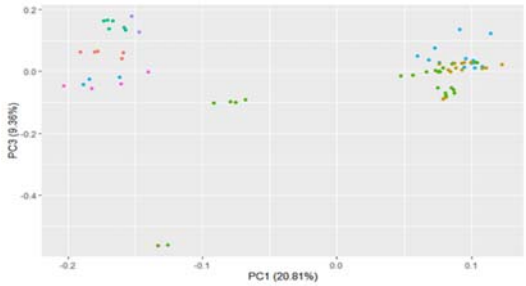


Figure 4: PCA plot of the dataset over PC1 and PC3. The two outliers are the green points in the lower part of the plot.

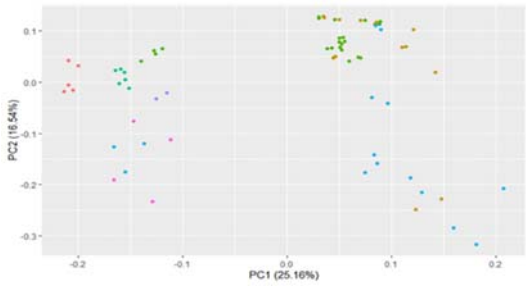


Figure 5: PCA plot of the dataset obtained after selecting a set of genes from Gene Ontology.

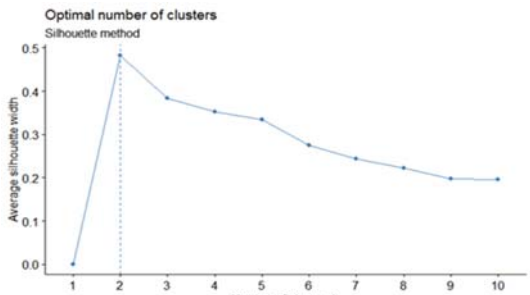


Figure 6: Plot of the silhouette method over different numbers of clusters. The dashed line corresponds to the optimal number of clusters.

The PCA plots in figure 3 have shown that the dataset may have some outliers. Plotting the dataset over the PC3, as shown in figure 4, reveals that two points determine much of the variability of this component, which explains 9.36% of the total. Therefore, we removed those two points, to avoid them from excessively affecting our results.

The dataset is under-sampled: it is made up of 70 samples of 24249 features. To reduce the number of features we selected a set of genes which may distinguish iPSC and ESC based on their biological function. In particular, we extracted from Gene Ontology[14] a list of 1518 genes that are involved in pathways related to embryogenesis and stemness. The PCA obtained keeping only these features (fig. 5) shows that most of the original variability has been retained.

To observe how the samples cluster in this new dataset we first chose the optimal number of clusters using the silhouette method (fig. 6), which shows that the optimal number of clusters is two. We plotted the corresponding silhouette

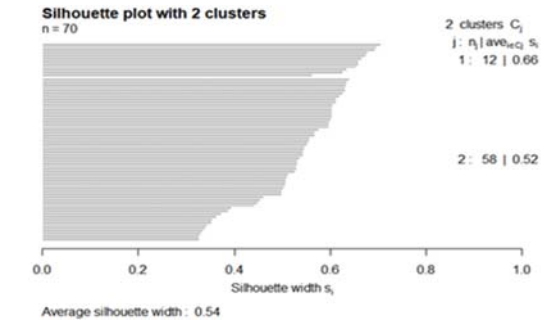


Figure 7: Silhouette of the dataset clustered in two groups

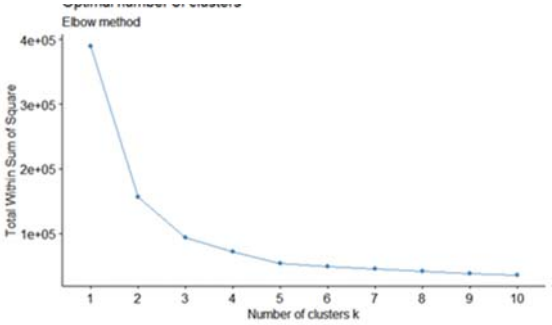


Figure 8: Plot of the within sum of squares over different numbers of clusters. We can observe an elbow around values 2 and 3.

to evaluate this clustering method in figure 7. The silhouettes show mostly positive values, which means the clusters are well defined. We also plotted the within-sum-of-squares in figure 8. It shows the presence of an elbow for k values around 2 or 3 which confirms the validity of our choice.

We computed a dendrogram using complete linkage highlighting the two clusters. Figures 9 and 10 show respectively the dendrograms obtained from both the original and the new dataset. Some similarities are noticeable, for example the same group of ESC seems to group separately from the other ones and the presence of two clusters is suggested in both dendrograms, however the one with features chosen from GO has a much clearer cluster separation.

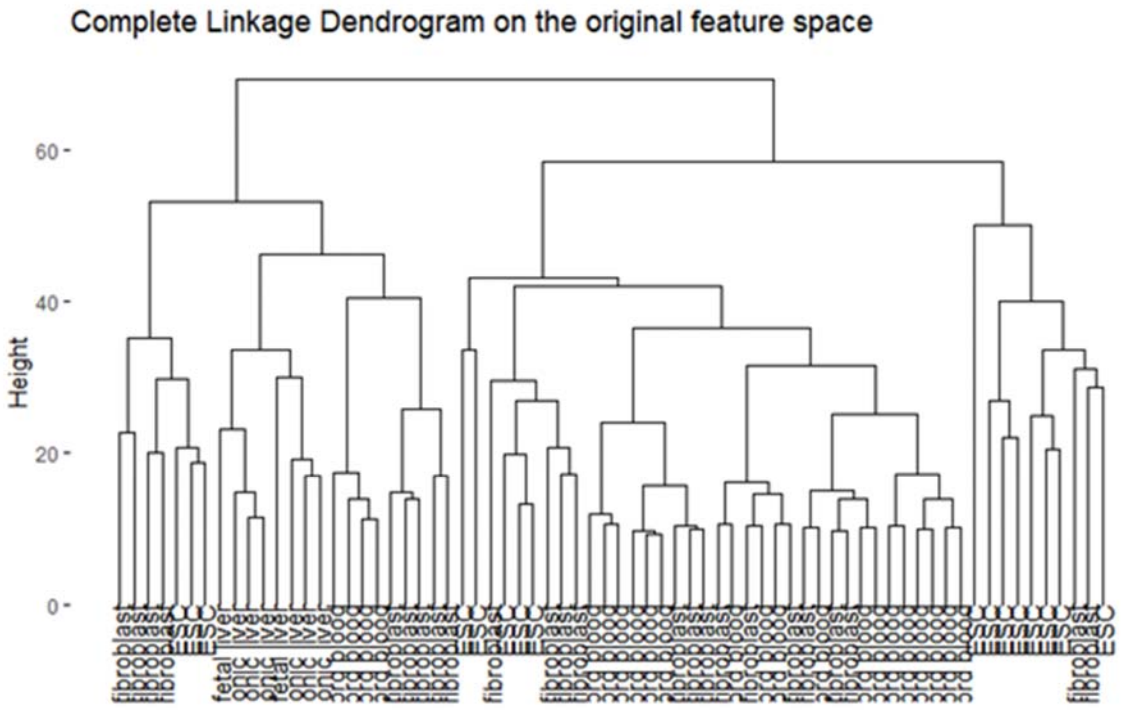


Figure 9: Complete linkage dendrogram of the GO dataset, split into two clusters.

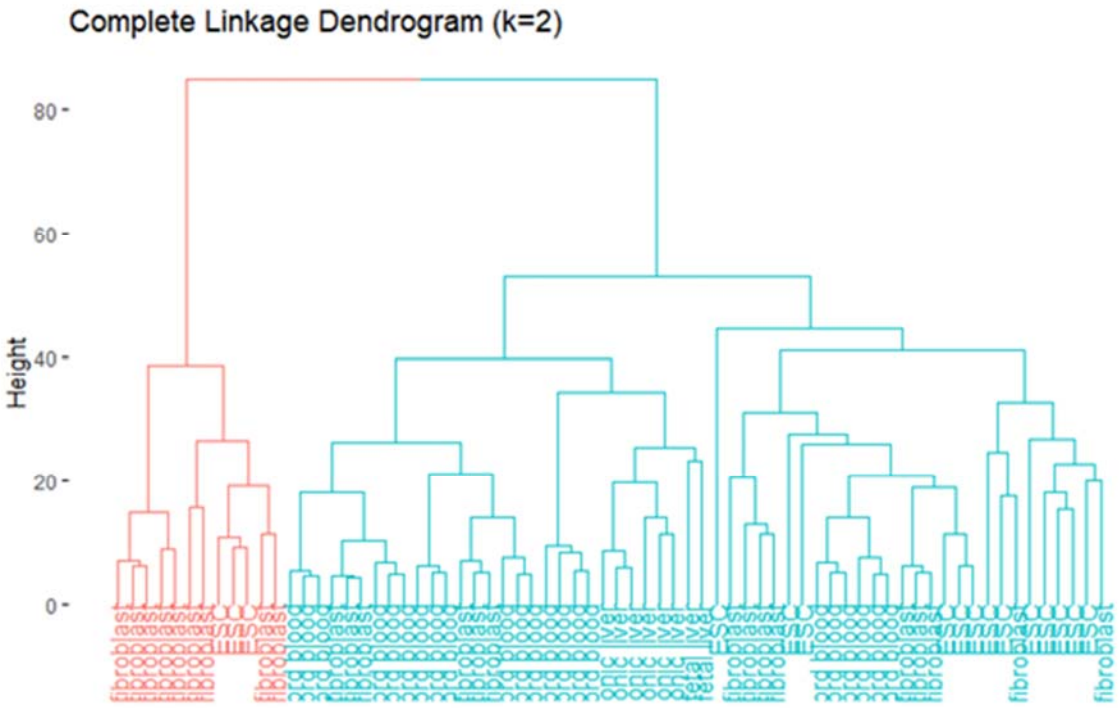


Figure 10: Complete linkage dendrogram of the original dataset, split into two clusters.

Results

Source tissue classification

There are 6 classes of source tissues in our dataset: "ESC", "cord blood", "adult fibroblast", "embryonic liver", "neonatal fibroblast", "AD specific fibroblast". In order to predict the tissue of origin from genetic expression we first employed kNN. We applied it to both a proxy dataset that we obtained from the most relevant components of PCA and to the GO dataset.

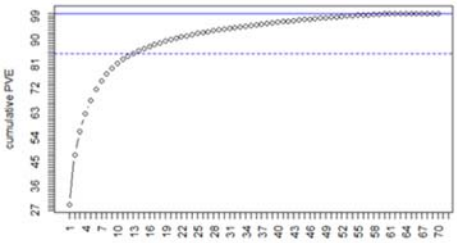


Figure 11: Cumulative PVE plot of the PCA on the dataset. The dashed line corresponds to the 85% variability threshold.

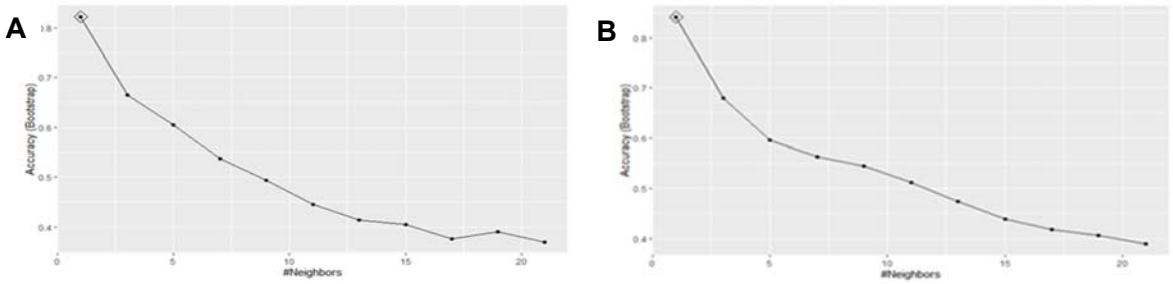


Figure 12: (A) Accuracy of the kNN models over different values of the parameter k on the proxy dataset. (B) Accuracy of the kNN models over different values of the parameter k on the GO dataset.

To create the proxy dataset, we first inspected the scree plot and the cumulative PVE plot (fig. 11). Lower variance PCs clearly expressed very little variability of the dataset. As such we believed that those components may be related to noise, so we used a restricted number of the first components to reduce feature space dimension and noise. We decided to preserve 85% of the variability, which corresponds to using the first 12 components. Then we evaluated the model over different values of k using bootstrap resampling. We found that the optimal k value is 1. Then we did the same on the GO dataset and still found that k=1 is the optimal parameter for kNN. Both accuracy plots are shown in figures 12A and 12B. Confusion matrices obtained by splitting the original datasets between test and train datasets for both methods are shown in figures 13A and 13B. The resulting accuracies are 87.5% for the proxy dataset and 93.75% for the GO dataset.

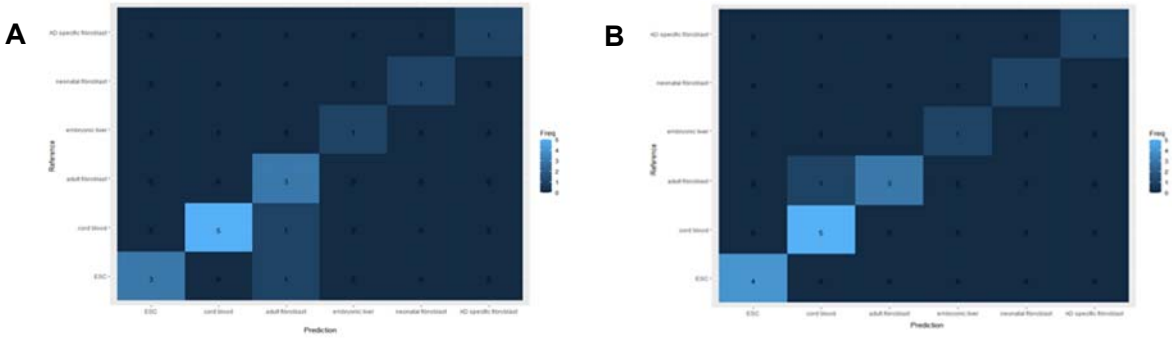


Figure 13: (A) Confusion matrix showing the performance of the kNN model on the proxy dataset. (B) Confusion matrix showing the performance of the kNN model on the GO dataset.

Gene-by-gene comparison of ESCs and IPSCs: differentially regulated genes (DEGs)

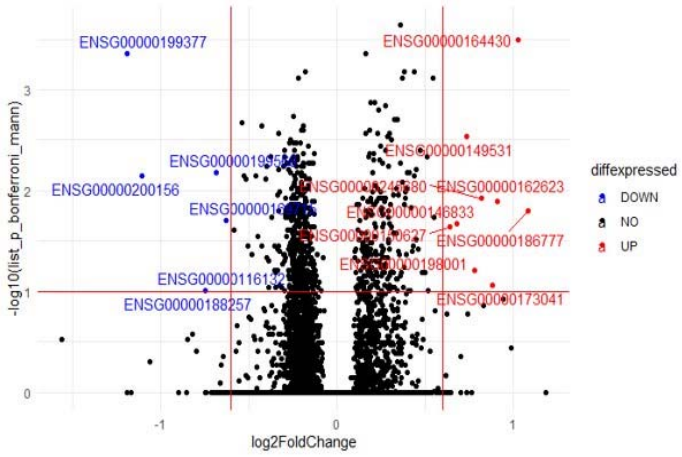


Figure 14: Volcano plot, each point is a gene. Log fold change is on the x-axis and the logarithm of corrected p-values on the y axis. The coloring is red, implying up-regulation in induced pluripotent stem cells with respect to the embryonic ones, or blue indicating the other way.

First of all, we wondered which genes are differentially regulated in embryonic versus induced pluripotent stem cells. By looking at the expression distributions and performing some normality checks, we found out that bimodality in some genes and heavily skewed distributions in others cannot even be rescued by data transformation. As a consequence, we recurred to the non-parametric alternative of the t-test, i.e., the Mann-Whitney U test. Since we were comparing seventy over-dispersed data points, our

samples, more than twenty-four thousand times, we needed to correct for multiple testing (we tried both the more conservative Bonferroni correction and the FDR). Moreover, the p-value only informs about the statistical significance, but the relative magnitude plays an equally important part in defining what differential expression is. This relative difference in expression is measured as the log fold change, indicating how many times a gene is more expressed in a group of samples with respect to the other one, on average, as shown in figure 14. Surely, by tuning the threshold on log fold change and p-value we can call “differentially expressed” more or fewer genes. As a start, we employed a strict constraint: Bonferroni correction with p-value < 0.10 and absolute value of the log fold change > 0.4. This screen identified 45 genes as DEGs.

Principal component analysis on the 45-DEGs space

The Mann-Whitney corrected p-values and the fold change provided us with a first screen on the genes which clearly discriminate embryonic versus induced pluripotent stem cells. By retaining only those genes as coordinates, hence describing our samples in a lower-dimensional space with respect to the original one, we expected a PCA on such space to discriminate well the embryonic and the induced pluripotent cells, and indeed it did, as shown in figure 15.

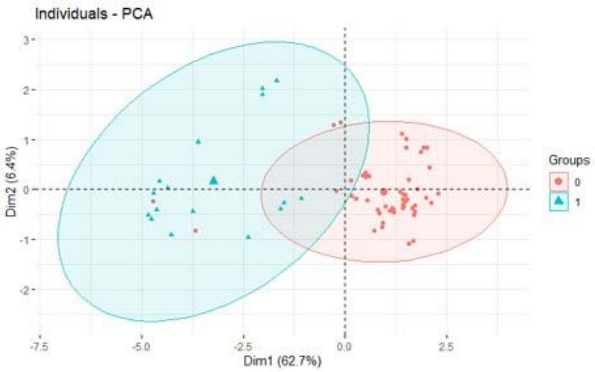


Figure 15: Principal Component Analysis (PC1 - Dim1 in plot, and PC2 - Dim2) calculated after expressing each sample in terms of the 45 strongly differentially expressed genes (DEGs). IPSCs in red, ESCs in blue.

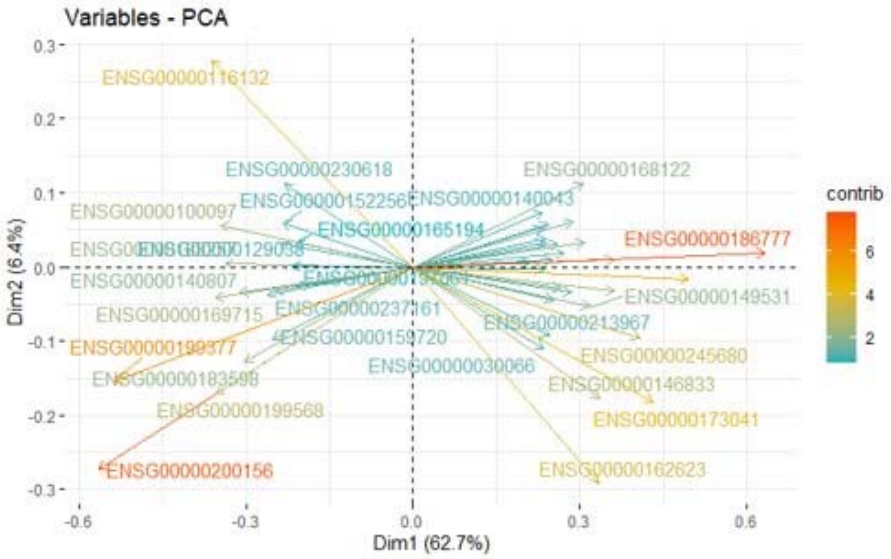


Figure 16: Genes contribution to the PC1 and PC2 of the 45-DEGs PCA. Each gene is in ENSG notation, and arrows coloring indicate the relative contribution (loading) to the first two PCs.

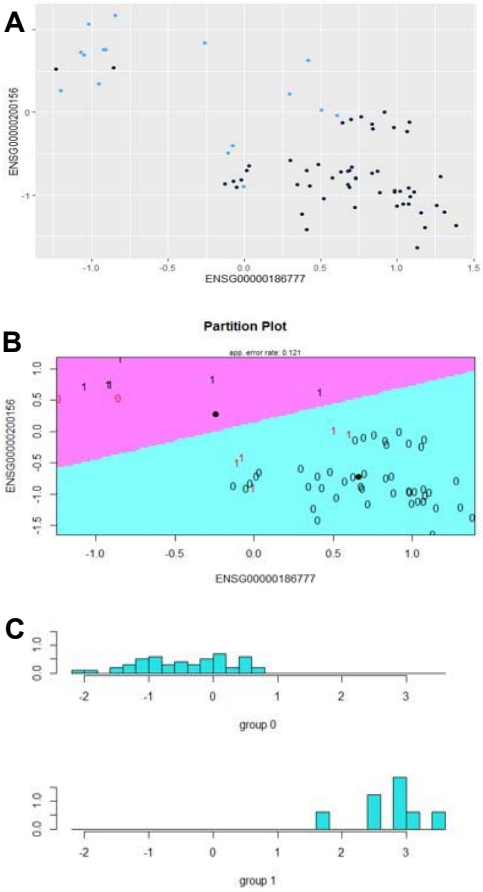


Figure 17: (A) Scatterplot of samples in the bidimensional space described by the two top-loader genes from the PCA on the 45-DEGs dataset. Dark blue corresponds to IPSCs, whereas ESCs are colored in light blue. (B) LDA partition plot on the bidimensional space in (A). (C) LDA histograms of the linear discriminant: IPSCs and ESCs are clearly separated, the border being around 1.

Performing this PCA also allowed us to pick the genes with higher loading, that is, which played a predominant role in the linear combinations of the first two principal components, ultimately separating our two clusters of cell phenotypes, as depicted in figure 16.

Two genes are enough to separate ESCs and iPSCs by LDA

Aiming to characterize relevant genes, to interpret the biology underlying our results, we identified the two top-loaders among the genes and we scattered our samples in the two-dimensional space described only by those genes (fig. 17). Embryonic stem cells (light blue) and induced pluripotent stem cells (dark blue) do separate well by looking only at this couple of genes. To confirm this visually compelling intuition, we performed linear discriminant analysis (LDA) on this bi-dimensional space. The partition plot and the histogram of the first and only discriminant are shown in figure 17. Accuracy is around 85% and slightly dependent on which samples end up in the training set. We are aware that LDA assumptions did not rigorously hold, but we trusted its results as it confirmed something we could actually see by eye, and deviations from assumptions are reasonably not too marked. We could not use a QDA as points were

Figure 18: EBI search for the top-loader gene ENSG00000186777 (ZNF732).

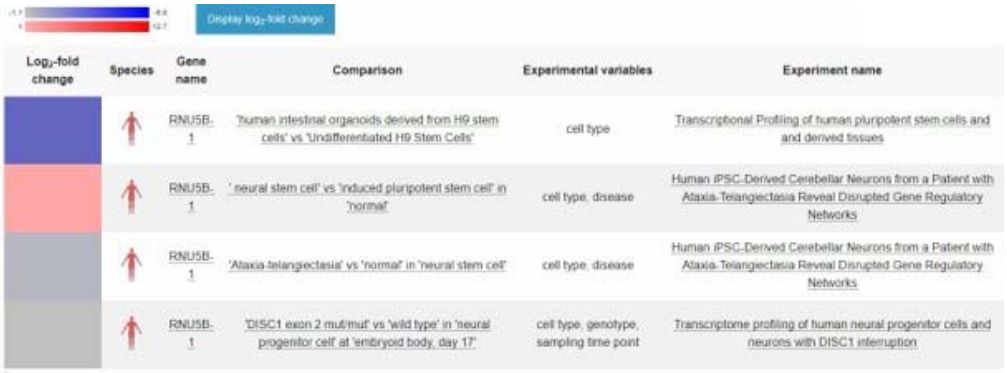


Figure 19: EBI search for the second top-loader gene ENSG00000200156 (RNU5B-1).

not enough – something to improve in our study was for certain the limited number of samples. The first gene identified is the transcription factor zinc finger protein 732. We confirmed on the EBI expression atlas[15] that this protein is really known to be someway related to induced pluripotency (fig. 18). The second gene produces an RNA implied in processing other RNAs, that is, a building block of the spliceosome, the U5B small nuclear 1 RNA. It also has something to do with stem cells, but it is not well characterized (fig.19).

Are there any better gene couples?

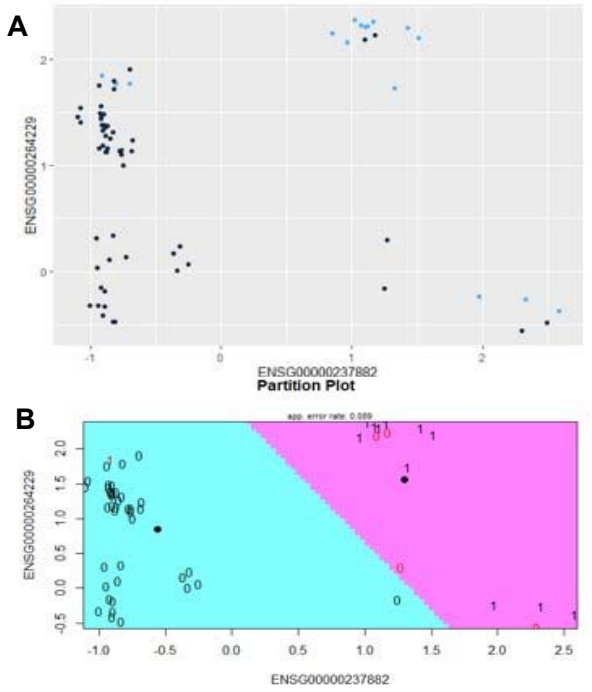
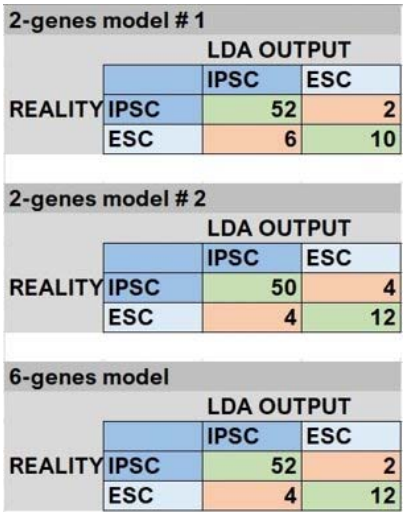


Figure 20: (A) Scatterplot of samples in the bidimensional space of the two top-loader genes in the analysis with no constraint on log-fold change, but only with FDR-corrected p-value < 0.01. Dark blue corresponds to IPSCs, whereas ESCs are colored in light blue. (B) LDA partition on the 2-genes plot in A.

The approach employed so far was maximally parsimonious: two genes were enough to differentiate embryonic and induced pluripotent stem cells with decent accuracy. We next wondered whether other gene couples could do a better or comparable job – we thus ran the previous pipeline but relaxing the cutoff for the log fold change, by starting with genes that had just a differential expression (FDR-corrected p-value<0.01, retaining 7020 genes). In fact, a different couple of genes emerged from this analysis, of which one is understudied and the second is, curiously, still implied in the spliceosome. By plotting our samples in the space of these two genes, however, one could see at a glance that iPSCs versus ESCs clustered more loosely and separated less precisely than before (fig. 20). A linear discriminant analysis nonetheless did not capture the cluster compactness or structure,



so the accuracy was comparable to that obtained with the first two genes. In any case, it is biologically plausible that embryonic and induced pluripotent cells differ substantially in more than two genes, so different couples of genes can yield comparable accuracy to discriminate between the two cell types (fig. 21). Sticking to this observation, we tried with different couples of genes, extracting them by chance among the differentially expressed, but the accuracy never reached the 85% of the previously rationally selected genes.

Figure 21: LOOCV classification scores in three models. (A) Samples coordinates are the two PCA top-loaders of the 45 top DEGs (ENSG00000186777, ENSG00000200156) (B) Samples coordinates are the two PCA top-loaders of the 7020 DEGs after relaxing log-fold change and p-value thresholds (ENSG00000237882, ENSG00000264229) (C) Samples coordinates are the six PCA top-loaders of the 7020 DEGs after relaxing log-fold change and p-value thresholds (ENSG00000237882, ENSG00000264229, ENSG00000201998, ENSG00000280434, ENSG00000239128, ENSG00000207344)

Less parsimonious models do not increase the classification accuracy

Aiming to increase the classification accuracy, we also tried less parsimonious models, by including all the 45 strongly differentially regulated genes, or the six top loaders, or the 19 top loaders (as here there was an elbow in the loading score). All the LDA models were thus subjected to leave-one-out cross-validation. Table 1 reports the three most accurate models. The 6-genes model yielded a 64 out of 70 accuracy, classifying correctly two more samples than the 2-genes models. Adding more genes did not yield any other gain in classification accuracy. It must also be noted how the performance on ESCs was always worse – but this was not a surprise as our dataset was strongly unbalanced in favor of iPSCs.

iPSC upregulated GO Terms

Mesenchyme migration and morphogenesis
Wnt signaling pathway
Heart development
Negative regulation of morphogenesis of an epithelium
Outflow tract septum morphogenesis
Chromatin organization involved in negative regulation of transcription
Negative regulation of gene expression, epigenetic
Endoderm development

Table 1: iPSC upregulated GO Terms - running GO main page with the 7020 top DEGs after relaxing log-fold change and p-value thresholds. The same analysis ran on the top 45 DEGs that retain information about the differential magnitude of expression does not lead to any meaningful information as those genes are too few.

Gene Ontology analysis of DEGs

Gene ontology has listed a number of human interpretable indications of biological processes associated with every studied gene. Basically, the user provides GO with a list of gene names, in our case the list of genes showing differential expression in ESCs and iPSCs, and the search engine determines a list of biological processes or cell components that appear to be affected[14,15]. Running this analysis, most emerging terms were expectedly tied to embryo development, including morphogenesis and migration, and to negative regulation of gene expression, possibly pointing at the temporary suppression of tissue-specific genes in the embryo.

Discussion

As induced pluripotent stem cells (iPSCs) are rapidly replacing embryonic stem cells (ESCs) in the literature, several groups have started comparing them, mostly thanks to the advent of multi-omics technologies[7]. In particular, cells were described as “RNA bags”, implying the transcriptome as a representative proxy of a cell type and state. To our knowledge, however, no one has ever tried to make a comprehensive analysis of all the ESCs and iPSCs published transcriptomes.

Our study represents one of the first efforts in this direction. However, lots of work should be done to overcome some limitations. First, the computational power at our disposal forced us to work only on the older and more sparse microarray data, instead of using scRNA-seq data. Secondly, our dataset suffers from under-sampling: the samples included come from manual curation, as we could not find a tool to comprehensively survey GEO to find iPSC and ESC experiments. Finally, and more importantly, we had to discard a lot of data as published microarray profiles are processed according to different platforms and algorithms with no trivial way to go back to the raw data and process them in a common framework. The problem of data format is nowadays well discussed by international consortia[16]. To overcome these issues all at once, we may employ powerful servers and work with the richest RNA-seq data. We are also considering writing a tool to automatically download and process these data from public repositories, finding keywords such as iPSC or ESC.

What we can already notice from our preliminary analysis is that the experiment and tissue of origin still differentiate iPSCs’ transcriptomes among them and to ESCs. Precisely, we found GO terms related to development to be up- or downregulated, pointing to intrinsic differences between the two macro-cell types, implying a need to better characterize whether these molecular differences are reflected into different behaviors that might compromise experiments or therapies if we use iPSC as they are today. Finding which gene modules are still differentially regulated between iPSCs and ESCs may guide experiments to treat iPSCs to manipulate specifically the expression of such modules.

Finally, our analytical framework is ready to be employed for more data-rich analysis, and it could be applied to different biological problems to dig deeper into transcriptomes to find the most parsimonious gene sets to discriminate between two (or more) cell types, which has a number of experimental implications, starting from in situ hybridization.

Appendix

All code and data is publicly available on Github:
https://github.com/lucafusarbassini/ipsc_vs_esc

- urls_complete.xlsx lists all the datasets about ESCs and IPSCs microarray data, manually curated from GEO
- data_extraction2.0.py automatically extracts the data from the aforementioned URLs and with specified datatype (depending on the algorithm for microarray data processing) and creates a dataset where each column is a sample and each row is a gene
- dataset.csv is the polished dataset, employed for all the analysis
- sample_to_info.csv lists useful information for the lookup table employed in the analysis below
- Script_R_2.1.Rmd is the code for data preparation, unsupervised analysis and 4 clusters classification
- ipsc_vs_esc.Rmd analyzes the transcriptomic differences between ESCs and IPSCs in a supervised fashion

All code is commented, unfortunately in Italian, as it was written for communication between the coders

References

1. Grealish, S. *et al.* Human ESC-Derived Dopamine Neurons Show Similar Preclinical Efficacy and Potency to Fetal Neurons when Grafted in a Rat Model of Parkinson's Disease. *Cell Stem Cell* **15**, 653–665 (2014).
2. Aoki, H. *et al.* Transplantation of cells from eye-like structures differentiated from embryonic stem cells in vitro and in vivo regeneration of retinal ganglion-like cells. *Graefe's Arch. Clin. Exp. Ophthalmol.* **246**, 255–265 (2008).
3. Blesch, A. Human ESC-Derived Interneurons Improve Major Consequences of Spinal Cord Injury. *Cell Stem Cell* **19**, 423–424 (2016).
4. Lo, B. & Parham, L. Ethical issues in stem cell research. *Endocr. Rev.* **30**, 204–213 (2009).
5. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).
6. Chin, M. H. *et al.* Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* **5**, 111–123 (2009).
7. Bilic, J. & Belmonte, J. C. I. Concise Review: Induced Pluripotent Stem Cells

Versus Embryonic Stem Cells: Close Enough or Yet Too Far Apart? *Stem Cells* **30**, 33–41 (2012).
8. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
9. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
10. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* **35**, W193–W200 (2007).
11. Birney, E. *et al.* An overview of Ensembl. *Genome Res.* **14**, 925–928 (2004).
12. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
13. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29–39 (2016).
14. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
15. Kapushesky, M. *et al.* Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res.* **38**, D690–D698 (2010).
16. Fujita, A., Sato, J. R., Rodrigues, L. de O., Ferreira, C. E. & Sogayar, M. C. Evaluating different methods of microarray data normalization. *BMC Bioinformatics* **7**, 469 (2006).