Supporting Regenerative Medicine by Integrative Dimensionality Reduction

F. Mulas[1]; L. Zagar[2]; B. Zupan[2,1,3]; R. Bellazzi[4,1]
[1]Centre for Tissue Engineering, University of Pavia, Pavia, Italy;
[2]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia;
[3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA;
[4]Dipartimento di Ingegneria Industriale e dell'Informazione, Università di Pavia, Pavia, Italy

Keywords
Stem cells, principal component analysis, gene subset selection, regenerative medicine, predictive modelling

Summary

Objective: The assessment of the developmental potential of stem cells is a crucial step towards their clinical application in regenerative medicine. It has been demonstrated that genome-wide expression profiles can predict the cellular differentiation stage by means of dimensionality reduction methods. Here we show that these techniques can be further strengthened to support decision making with i) a novel strategy for gene selection; ii) methods for combining the evidence from multiple data sets.

Methods: We propose to exploit dimensionality reduction methods for the selection of genes specifically activated in different stages of differentiation. To obtain an integrated predictive model, the expression values of the selected genes from multiple data sets are combined. We investigated distinct approaches that either aggregate data sets or use learning ensembles.

Results: We analyzed the performance of the proposed methods on six publicly available data sets. The selection procedure identified a reduced subset of genes whose expression values gave rise to an accurate stage prediction. The assessment of predictive accuracy demonstrated a high quality of predictions for most of the data integration methods presented.

Conclusion: The experimental results highlighted the main potentials of proposed approaches. These include the ability to predict the true staging by combining multiple training data sets when this could not be inferred from a single data source, and to focus the analysis on a reduced list of genes of similar predictive performance.

Correspondence to:
Riccardo Bellazzi
Centre for Tissue Engineering
University of Pavia
Pavia
Italy
E-mail: riccardo.bellazzi@unipv.it

Methods Inf Med 2012; 51: 341–347
doi: 10.3414/ME11-02-0045
received: November 8, 2011
accepted: May 4, 2012
prepublished: July 5, 2012

1. Introduction

Stem cells are self-renewing populations of cells that can give rise to diverse specialized cell types. In mammals, embryonic stem cells (ESCs) can be isolated, proliferated and differentiated in vitro into a potentially unlimited variety of tissues. The same pluripotent capability is attributed to induced pluripotent stem cells (iPSCs), reprogrammed adult cells obtained by inducing the expression of specific genes [1]. Some types of stem cells have been considered for use in gene therapy [2, 3], i.e. inserting normal genes into a person's cells for restoring the correct functioning of tissues and organs affected by a particular genetic disorder [4]. Thanks to their self-renewal properties, stem cells would eliminate the need to provide repeated administrations of the therapy. The use of iPSCs is considered appropriate for gene therapy, as the cells can be generated from the individual's own tissues. These patient-specific healthy cells can be transplanted, avoiding problems with rejection. During this process, cells may fail from successful reprogramming or differentiating and may remain trapped into partially differentiated states due to several factors [5]. An accurate monitoring of the pluripotency level is thus required to make iPSCs transplantation a safe and efficient practice in regenerative medicine.

Keeping under control the molecular signature that characterizes cellular differentiation is particularly hard. At present, the standard assay for pluripotency of stem cells is the generation of different types of tumours in immunodeficient mice [6]. Alternative methods for assessment of cellular developmental potency have recently gained interest. Given the increasing use of genomic information for clinical practice [7, 8], different experiments have been carried out to derive a pluripotency signature from microarray-based gene expression data [9–11]. These studies have shown that a suitable approach to transform the whole-genome transcriptome profiles into a predictive model of cell differentiation is to apply dimensionality reduction techniques to the RNA samples from various stages of development. Recently, Muller et al. [11] used a non-negative matrix factorization algorithm to represent stem cell data in a low-dimensional space. This representation allowed testing two classification methods that distinguish pluripotent from non-pluripotent samples. Aiba et

al. [10] studied the developmental potency of differentiating mouse embryonic stem cells and showed that the positions of the samples along trajectories presented in 3-dimensional space reflect the developmental potency of the cells.

In a recent work [12], we adopted an approach similar to the one proposed by Aiba et al. to predict the differentiation stage of a cell from its genome-wide transcription profile. We compared several methods that transform a sample expression profile into a real-valued projection. A one-dimensional ruler, which we refer to as differentiation scale was obtained by mapping the projections of differentiation samples to a one dimensional space. Samples from new experimental settings (e.g. iPSCs or cellular lines treated with chemical agents) were projected on a scale developed in standard conditions to uncover the actual stage of development with respect to the normal dynamics of differentiation. The predictive accuracy of the proposed methods was assessed computationally, either by cross-validation within the same data set or by training a predictive model on one experiment and testing the predictions on another one. We have examined both methods that construct the projections explicitly, such as Principal Component Analysis (PCA), and methods that can rank the presented samples, such as Minimum Curvilinear Embedding [9]. We have also observed that these methods in combination with the computationally selected 1000 best-ranked genes ensure a high quality of stage prediction, as opposed to using only known pluripotency markers. Among the different data transformation and ranking approaches tested we preferred the PCA because of its simplicity and the benefits of providing an explicit prediction model. Although the aim of our work was similar to the one studied by Muller et al., the PCA-inferred model does not only assign a classification label to characterize the pluripotency of a new sample. In addition, the differentiation scale also provides for visualization of the placement of the sample on a scale where the distances reflect the differences among samples on a phenotypic level.

The application of dimensionality reduction techniques to the analysis of stem cell data started to be explored only recently. One of its potential gains is to improve the robustness of the predictions and the consistency of the proposed methods in order to develop a reliable decision-making tool useful in the field of regenerative medicine. With the increasing number of available stem cell data sets in public repositories, one of the crucial goals in this direction is to derive a robust differentiation scale based on the evidence from different experiments with similar observed processes and experimental settings. This integrative and reliable tool should be coupled with a selection of those genes that are actually responsible for determining the pluripotency signature of each stage. To this end, it is straightforward to ask if the differentiation scale could benefit from a stage-specific gene selection as well as how much a low number of key genes impacts the performance of our integrative predictive device.

In this report we show how the PCA-inferred differentiation scale model can be enhanced with i) a novel strategy for selecting the genes used by the model and ii) methods for combining the evidence from multiple training data sets. A summary of the proposed methods is shown in ▶Figure 1.

We here report on some major advances with respect to our previous approach. First, we propose a new strategy for gene selection that specifically characterize a particular stage of development. The information of which genes can be used as transcriptional markers of each developmental stage is valuable for researchers in the area of regenerative medicine. Selected genes may be used to investigate the biological processes and pathways activated in specific phases along differentiation. Second, the goal of this work is to provide a robust differentiation scale by inferring the integrated model from a collection of data sets. To this aim, a number of methods for combining gene expression data sets are evaluated. Finally, we also demonstrate that reduced lists of genes can be successfully used in the integrative model, providing high quality stage prediction with reduced computational costs.

The paper is structured as follows: in Section 2 we introduce the PCA-based differentiation scale model and we present the approaches used for extraction of the most informative genes and integration of a collection of data sets. In section 3 we show the performances of the algorithms on six data

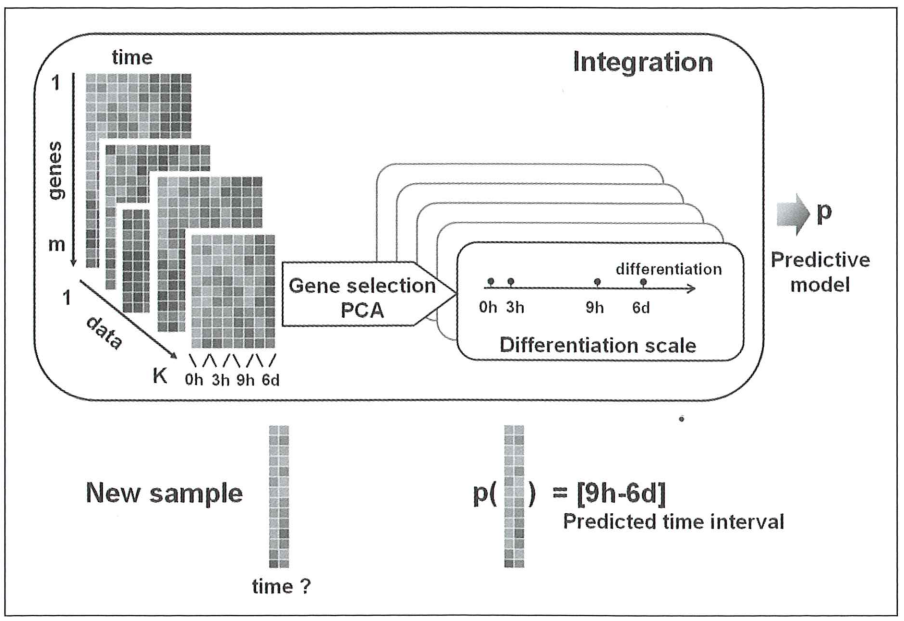

Fig. 1 Differentiation stage prediction from a collection of data sets. A differentiation scale with a position for each stage can be obtained from one data set by applying gene selection strategies and principal component analysis. Integration methods combine multiple experiments in standard culturing conditions to infer a model that predicts the developmental stage of a new uncharacterized sample.

sets from NCBI's Gene Expression Omnibus and application to iPSCs data. In Section 4 we present a discussion of the main features of the algorithms. Finally, Section 5 presents some concluding remarks and future directions of the work.

2. Methods

To infer a stage prediction model and its graphical representation, we apply PCA to a data set containing genome-wide expression measurements for m genes in n different samples along differentiation. The data can be represented with an $n \times m$ matrix D containing expression values d_{ji} for each gene i measured in sample j. Using PCA, a real number $p(d_j)$ can be assigned to sample j by projecting its expression profile $d_j = (d_{j1}, \ldots, d_{jm})$ to the first principal component:

$$p(d_j) = \sum_{i=1}^{m} d_{ji} w_i \qquad (1)$$

with w_i being the elements of the first eigenvector of the covariance matrix $D^T D$, for a mean-centred data D. Each sample is accompanied by the information about the stage, that is, its ordering in the development process. Such information can be obtained from the development time at which the measurement was performed (e.g. "3 h" or "6 d", as shown in ▶Fig. 1). When multiple samples from the same stage are included in the data we use the median of their PCA projections. The final result is a set of real numbers (projections of stages) that can be used to construct a one-dimensional ruler – the differentiation scale.

A gene score GS_{is} that reflects the importance of gene i in the stage of development s can be derived using the PCA-inferred weight w_i and the average expression value of gene i in the samples from stage s:

$$GS_{is} = w_i a_{is}, \; a_{is} = \frac{1}{|s|} \sum_{j \in s} d_{ji} \qquad (2)$$

Since weights w_i are real numbers, gene score GS_{is} can be either positive or negative. In the following we show how this score is used to select the most informative genes. Other gene subset selection methodologies and methods for prediction based on multiple data sets are also described.

2.1 Gene Selection

The first step of our analysis is the selection of the genes whose expression values are used to infer the predictive model. A fold-change approach for time series data [13], we referred to as *FC*, was tested in our previous work. We ranked the genes based on the number of time points where the fold change value with respect to the gene expression at the initial time point was at least 2-fold. A similar approach, named *DIFF*, ranks the genes based on the difference between the average of expression values of samples from the last and from the first time point [14]. Finally, we used *AREA*, a method that ranks the genes based on the area bounded by the gene expression time series and a constant profile with value equal to the expression at the first time point [15].

In our previous work, we observed high quality of prediction when a set of 1000 best-ranked genes was used for model training. In this work, we compared these methods with a strategy that selects a reduced number of genes for prediction, that we refer to as *Stage-Specific Filtering (SSF)*. For each stage s, *SSF* extracts genes with the highest scores GS_{is} (▶Eq. 2). The lowest possible number of genes is extracted for which the sum of their scores is more than a chosen proportion (e.g. 99%) of the total sum. Genes selected in at least one stage were used as features for the predictive model.

2.2 Integration of Different Data Sets

To address the problem of data integration in the context of stage prediction we have tested different approaches. The first one we will refer to as Merging is based on the aggregation of samples from separate experiments. Numerous efforts have been made to demonstrate the technical equivalence of microarray data across experiments, when proper normalization methods are used to handle the variability that characterizes different laboratories and experimental protocols [14, 16, 17]. Several works directly concatenated the gene expression values from different samples, or developed a meta-analysis to obtain an integrated score for each gene [18]. In this study, we relied on a pipeline for combining microarray data from NCBI GEO proposed by Dudley et al. [14]. For each data set in our collection, we applied quantile normalization [17, 19] and we collapsed the probes by computing the mean value of all probes with the same Entrez Gene identifier. Afterwards, the arrays were merged. Each gene in

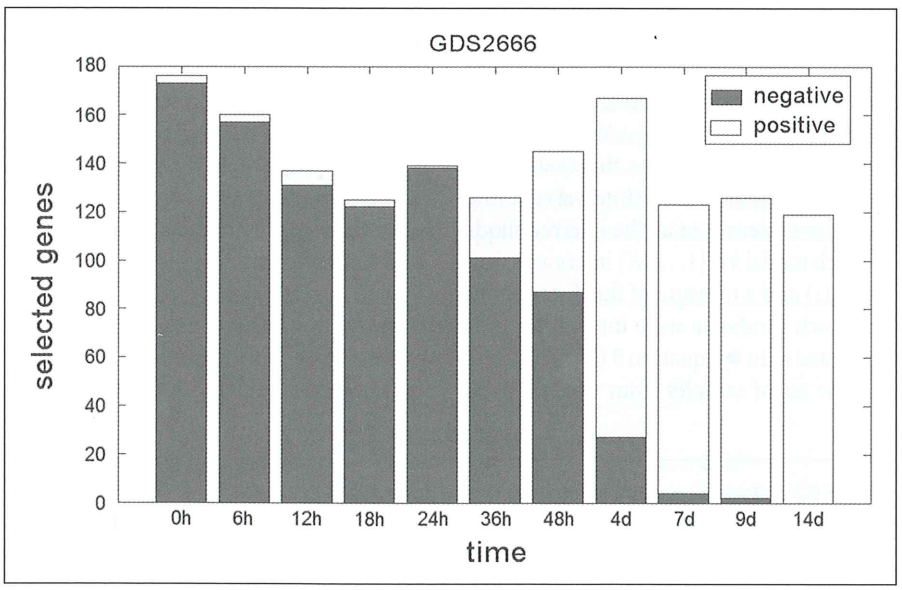

Fig. 2 Number of genes selected with *SSF* (α = 99%) in each stage for data set GDS2666

the data set collection was described with an expression value for each stage, obtained by computing the mean value across all samples measured at the same time point. The result of applying PCA on the expression values of the most informative genes and thus selecting the first PCA is a differentiation scale that combines all the K data sets. New samples can be projected on the integrated scale to predict the developmental stage. A variant of this method called *Merging ranks* is obtained by applying rank normalization, i.e. replacing each gene expression value by its rank in the considered sample [14, 20]. The average of the rank values across all the samples from the same time point is then considered to construct an aggregate data set.

Rank normalization was applied to each data set in our collection and used to develop a third method, called *Ranks*. In this approach we created an integrated matrix containing only genes measured in all the data sets. Instead of merging the arrays, we here considered separately the rank values from all the samples measured in each data set.

Another proposed method, here referred to as *Voting*, borrows the main idea from ensemble classification algorithms. These methods have been shown to improve the accuracy of classifiers [21, 22]. Given a training set, the voting algorithms infer different classifiers either by sampling from the data or running different classification techniques. The results are combined to create a final classifier that predicts the stage with the most votes. Our algorithm consists of an ensemble of PCA models. A PCA model is trained on each data set, obtaining K different predictive models – differentiation scales. Given a new sample s, the result of this method is a predicted time interval resulting from the consensus of all the inferred models. Each model $k \in \{1, \ldots, K\}$ infers a projection $p_k(s)$ and a measure of the distance of s from each candidate stage interval $[t_i, t_j]$ is computed as in ▶ Equation 3 (Fig. 3), where S_i is the set of samples from time point t_i, $p_k(x)$ is a 1-D projection of sample x inferred with the k-th model, Me(*set*) the median of the values in *set*. The time intervals are ranked by each model based on the distances $D_{k,ij}$. For each interval, the sum of the ranks from all models is then used as a score and the best scoring interval, i.e. with the minimum sum of ranks is returned as a prediction.

As an alternative approach, we applied Two-Dimensional Principal Component Analysis (*2DPCA*) [23]. Within this method, the optimal projection of a set of matrices $\{D_i\}$, $i = 1, 2, \ldots, K$ is given by the eigenvector corresponding to the largest eigenvalue of the covariance matrix $S_D = \frac{1}{K}\sum_{i=1}^{K}[D_i - \bar{D}]^T[D_i - \bar{D}]$, $\bar{D}$ being the average matrix of all training data sets. These matrices must be of the same dimension, so we reduced all the data sets to a common set of probes and constructed a set of matrices, one for each time point, containing the average expression values of the common genes in all the data sets. The resulting covariance matrix was used to obtain the projections for test samples.

2.3 Evaluation

We have compared the techniques for construction of integrated differentiation scale described in Section 2.2, and combined the best integration model with the methods for gene selection presented in Section 2.1. The following procedure was applied:

1. Infer a differentiation scale model from a collection of training data sets.
2. Obtain projections for samples from a test data set.
3. Score the quality of the stage predictions of test samples.

In the training phase, the model learns PCA weights from the integrated data sets, or separately from each training data set in the case of *Voting* method. These weights are used to combine the gene expression values and construct a differentiation scale, or separate scales for *Voting*.

Given a test sample t, the integrative model predicts its differentiation stage in the form of a projection on the scale, which should correspond to its real stage of development. For instance, if a sample from normal experimental conditions has been collected after four hours along differentiation, the projected data point should be placed between "3h" and "9h" on the scale depicted in ▶ Figure 1, i.e. $p(t) > p(0h)$, $p(t) > p(3h)$, $p(t) < p(9h)$ and $p(t) < p(6d)$. Formally, we have compared each test sample s_x collected at time point t_x with each training sample s_y from time point t_y, and we checked the ranking of their projections. For *Voting* method, the ranking predicted by the majority of the training data sets was considered. In standard conditions, the predicted ranking should preserve the original order of samples. i.e. $p(s_x) > p(s_y) \Leftrightarrow t_x > t_y$. We checked this condition for every pair of samples from two different stages of development. The proportion of pairs for which the original ranking was preserved was taken as a quantitative measure of the quality of predictions. This measure corresponds to the *C*-score, a generalization of the area under receiver operating characteristic curve [24].

3. Results

In this section we show the evaluation and the application of two of the proposed methods to a set of stem cell data from NCBI's Gene Expression Omnibus [25]. We focused on six data sets (GDS2666, GDS2667, GDS2668, GDS2669, GDS2671, GDS2672) from the same study that analyze differentiation into embryoid bodies for three distinct but genetically similar mouse ESC lines. We also show an application of the integrative tool to predict the pluripotency status of iPSCs.

3.1 Predictive Accuracy of the Integrative Tool

In the following, the results in terms of predictive accuracy of integration methods

$$D_{k,ij} = \frac{|p_k(s) - \mathrm{Me}(\{p_k(x): x \in S_i\})| + |p_k(s) - \mathrm{Me}(\{p_k(y): y \in S_j\})|}{2}, \quad j = i + 1$$

Fig. 3 Distance of a sample s from the $[t_i, t_j]$ time interval

and the contribution of SSF and other gene selection methods are presented. ▶Figure 2 shows an example of stage-specific gene selection on one data set. In general, we noticed that the number of genes that actually determine the position of each point on the scale was considerably lower than 1000. Moreover, the proportion of positive terms that contribute to the projection increases along differentiation, thus enabling the construction of a scale where, in normal conditions, the true order of the projections is preserved.

We have first evaluated the performance of all the methods for integration presented in Section 2.2, regardless of gene selection. ▶Table 1 shows the *C*-scores obtained when one data set is used as a test and the other five data sets are combined with the methods for integration previously described. All the methods showed high quality of prediction, with Merging obtaining the best performance. To highlight the statistical differences among the tested methods, we ranked them for each data set separately and we compared their average ranks. Significant differences were observed with Bonferroni-Dunn post hoc test [26] ($p < 0.05$) between *Merging* and both *Ranks* and *2DPCA* methods. The constraint of using genes measured in all the training data sets (about 3000 genes) required from these two methods did not provide sufficient information for prediction.

In a second step, we added different gene selection methods to the best scoring integration strategy, *Merging*. The total number of genes selected with *SSF* (α = 99.5%) on the merged data sets ranged from 239 to a maximum of 292 when GDS2668 and GDS2666 were used as test data sets, respectively. The aim of this analysis was to compare different methods for the selection of a reduced set of genes, as well as to investigate the contribution of integration and gene selection to the predictive model. The performance of our proposed method, *SSF*, was compared with other gene ranking methods when the same number of genes selected by *SSF* was used. The combined merging and selection methods were compared with the approaches without gene selection (*Merging*) and without integration (*SSF*). In the last case we considered the average of the scores obtained for each test data set when the training data were used separately for prediction. All the scores obtained with the methods that include gene selection were around 0.8 with average scores ranging from 0.895 for *Merging-SSF* to 0.844 for *SSF*. The statistical analysis of the methods' performance ranks is summarized in ▶Figure 4. Integration methods performed better than the approach without integration (*SSF*), with a significant difference for the two best-ranked methods. *Merging-SSF* was not significantly different from the best method (*Merging*). We thus prefer the variant with a reduced number of genes and we used *Merging-SSF* to predict the developmental stage of reprogrammed cells.

Table 1 C-scores on six data sets from GEO. The average score and the average rank for each test data set is reported.

Test data	Merging	Voting	Merging ranks	Ranks	2DPCA
GDS2666	0.977	0.954	0.940	0.491	0.737
GDS2667	0.953	0.909	0.935	0.514	0.753
GDS2668	0.900	0.936	0.858	0.498	0.794
GDS2669	0.900	0.839	0.919	0.523	0.776
GDS2671	0.867	0.830	0.878	0.534	0.742
GDS2672	0.864	0.875	0.862	0.509	0.622
Avg. score	**0.910**	**0.891**	**0.899**	**0.512**	**0.738**
Avg. rank	**1.667**	**2.167**	**2.167**	**5.000**	**4.000**

3.2 Application

To test the utility of our integrative device, we applied *Merging-SSF* to predict the developmental stage of pluripotent, partially reprogrammed and differentiated mouse cell lines [27]. These cells were studied by Mikkelsen et al. [27] to characterize the genomic signature underlying various stages of the reprogramming process. Reprogrammed mouse embryonic fibroblasts (r-MEFs) were obtained after 16 days of culture with reprogramming factors, and only the 1,2% of cells achieved complete reprogramming to a pluripotent state. The obtained fully reprogrammed cells (iPSCs) and cell lines trapped in a partially reprogrammed state (MCV8) were isolated from the culture. A small portion of the latter was found positive for the stem cell marker SSEA1 (MCV8+), and analyzed separately from the less pluripotent SSEA1-negative cells (MCV8–).

First, the six Gene Expression Omnibus data sets were used to infer six separate scales onto which we have projected the data from four reprogrammed fibroblast samples (▶Fig. 5a). Next, Merging-SSF was applied. The projections on the integrated

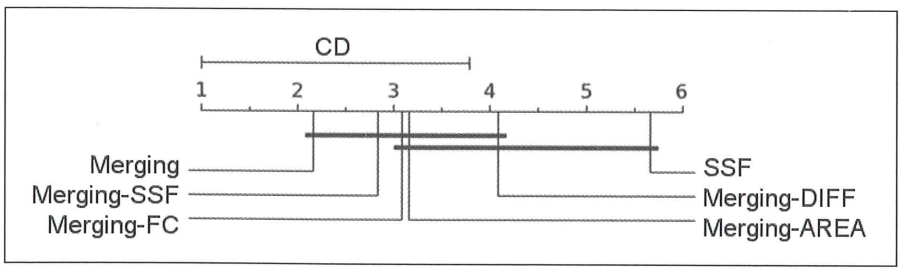

Fig. 4 Statistical analysis of rank differences. Critical distance (CD) groups methods whose performances ranks are not statistically different according to Bonferroni-Dunn post hoc test ($p < 0.05$)

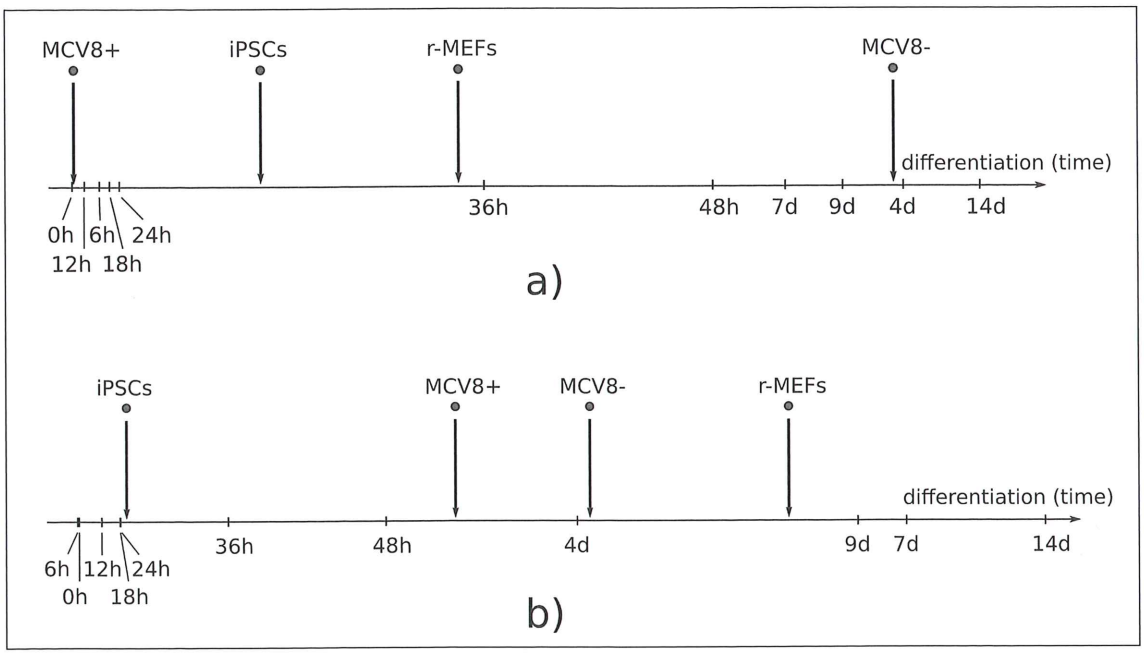

Fig. 5
Differentiation scale and projection of reprogrammed fibroblast samples to the scales developed on a) one data set (GDS2669) and b) integration of the six data sets with Merging-SSF

scale obtained from all the six data sets with Merging-SSF method (▶Fig. 5b) confirmed the results of the genomic analysis. In fact, iPSCs were projected next to early stages, indicating a pluripotency level highly similar to undifferentiated embryonic stem cells. On the contrary, MCV8+ and MCV8– were predicted as more differentiated samples, with MCV8+ associated to a more pluripotent status than MCV8–. The projection of r-MEFs is also coherent with the characterization of these cells, where the majority of cells was prevented from reaching a de-differentiated state, resulting in a genomic profile similar to partially reprogrammed cell lines. As shown in ▶Figure 5, for some data sets the inferred ordering of the projections did not correspond to the real ordering of samples as described by Mikkelsen et al. and confirmed with *Merging-SSF* model.

4. Discussion

The paper describes methods to deal with several data sets in order to derive reliable models for the prediction of differentiation stage of cells. Besides standard methodologies for selecting genes that exhibit a changed profile over time, we explored a method that exploits PCA eigenvectors to weigh genes for their contribution to the differentiation potential in each stage. Selected genes were used as features for integrative models that combine data from several experiments. We observed remarkable differences among some of the tested integration methods in terms of predictive accuracy. *Merging*, *Voting* and *Merging* ranks performed very well. Despite that, some differences that have impact on the application of those methods should be underlined. *Merging* has the advantage of providing a common differentiation scale, which illustrates the dynamics of the process through a graphical representation. On the other hand, the data sets considered for integration should refer to similar experimental settings, since this approach directly aggregates all the samples by computing average values for the included genes. *Voting* is independent from the platform and the experimental protocols, since separate models are developed. It does not result in a common scale but the different scales are used to predict a consensus-based time interval. Surprisingly, a small number of genes selected with *SSF* strategy performed very well and similarly to the approach that uses all the genes included in the microarray profile. At this stage of evaluation, we preferred the *Merging* integration method accompanied with the *SSF* gene selection strategy and we applied it successfully to predict differentiation stages of reprogrammed cells.

5. Conclusion

Stem cells and iPSCs offer exciting promise for personalized therapies in regenerative medicine, but several analyses have to be carried out for monitoring the real differentiation stage of cells before transplantation. We have presented and evaluated methods for the creation of a decision support tool that aims to become a reliable prediction instrument of differentiation stage. After the evaluation of different approaches, we identified a data integration method, *Merging*, that enables accurate and robust predictions. To complete the integrative tool, we added a gene selection strategy, *SSF*, which provide insights into the differentiation potential of each stage and is used to select the features of the integrative model, obtaining good performances. The results highlighted the capability of the integrative tool, *Merging-SSF*, to predict the correct order of samples in a collection of data sets, and predicts well also in cases when the real order is difficult to be inferred from a single model-differentiation scale. The utility of the proposed methods needs to be further confirmed on a larger collection of data sets, once more

experiments on stem cell differentiation in standard conditions will become available.

Acknowledgements

This work was supported by the Fondazione Cariplo grant (2008–2006) "Bioinformatics for Tissue Engineering: Creation of an International Research Group" and by EU FP7 project "CARE-MI" (Health-F5-2010-242038). LZ and BZ were also supported by grants from Slovenian Research Agency (P2-0209, J2-9699, L2-1112). Alberto Averna is gratefully acknowledged for his help in software development.

References

1. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 2006; 126 (4): 663–676.
2. Cartier N, Hacein-Bey-Abina S, Bartholomae CC, Veres G, Schmidt M, Kutschera I, Vidaud M, Abel U, Dal-Cortivo L, Caccavelli L, Mahlaoui N, Kiermer V, Mittelstaedt D, Bellesme C, Lahlou N, Lefrere F, Blanche S, Audit M, Payen E, Leboulch P, l'Homme B, Bougneres P, Von Kalle C, Fischer A, Cavazzana-Calvo M, Aubourg P. Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. Science 2009; 326 (5954): 818–823.
3. Sun XY, Nong J, Qin K, Warnock GL, Dai LJ. Mesenchymal stem cell-mediated cancer therapy: A dual-targeted strategy of personalized medicine. World J Stem Cells 2011; 3 (11): 96–103.
4. Giordano FA, Hotz-Wagenblatt A, Lauterborn D, Appelt JU, Fellenberg K, Nagy KZ, Zeller WJ, Suhai S, Fruehauf S, Laufs S. New bioinformatic strategies to rapidly characterize retroviral integration sites of gene therapy vectors. Methods Inf Med 2007; 46 (5): 542–547.
5. Okita K, Yamanaka S. Induced pluripotent stem cells: opportunities and challenges. Philos Trans R Soc Lond B Biol Sci 2011; 366 (1575): 2198–2207.
6. Daley GQ, Lensch MW, Jaenisch R, Meissner A, Plath K, Yamanaka S. Broader implications of defining standards for the pluripotency of iPSCs. Cell Stem Cell 2009; 4 (3): 200–1; author reply 02.
7. Maojo V, Martin-Sanchez F. Bioinformatics: towards new directions for public health. Methods Inf Med 2004; 43 (3): 208–214.
8. Bicciato S, Luchini A, Di Bello C. Marker identification and classification of cancer types using gene expression data and SIMCA. Methods Inf Med 2004; 43 (1): 4–8.
9. Cannistraci CV, Ravasi T, Montevecchi FM, Ideker T, Alessio M. Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes. Bioinformatics 2010; 26 (18): i531–9.
10. Aiba K, Nedorezov T, Piao Y, Nishiyama A, Matoba R, Sharova LV, Sharov AA, Yamanaka S, Niwa H, Ko MS. Defining developmental potency and cell lineage trajectories by expression profiling of differentiating mouse embryonic stem cells. DNA Res 2009; 16 (1): 73–80.
11. Muller FJ, Schuldt BM, Williams R, Mason D, Altun G, Papapetrou EP, Danner S, Goldmann JE, Herbst A, Schmidt NO, Aldenhoff JB, Laurent LC, Loring JF. A bioinformatic assay for pluripotency in human cells. Nat Methods 2011; 8 (4): 315–317.
12. Zagar L, Mulas F, Garagna S, Zuccotti M, Bellazzi R, Zupan B. Stage prediction of embryonic stem cell differentiation from genome-wide expression data. Bioinformatics 2011; 27 (18): 2546–2553.
13. Park T, Yi SG, Lee S, Lee SY, Yoo DH, Ahn JI, Lee YS. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. Bioinformatics 2003; 19 (6): 694–703.
14. Dudley JT, Tibshirani R, Deshpande T, Butte AJ. Disease signatures are robust across tissues and experiments. Mol Syst Biol 2009; 5 307.
15. Di Camillo B, Toffolo G, Nair SK, Greenlund LJ, Cobelli C. Significance analysis of microarray transcript levels in time series experiments. BMC Bioinformatics 2007; 8 (Suppl 1): S10.
16. Severgnini M, Bicciato S, Mangano E, Scarlatti F, Mezzelani A, Mattioli M, Ghidoni R, Peano C, Bonnal R, Viti F, Milanesi L, De Bellis G, Battaglia C. Strategies for comparing gene expression profiles from different microarray platforms: application to a case-control experiment. Anal Biochem 2006; 353 (1): 43–56.
17. Boes T, Neuhauser M. Normalization for Affymetrix GeneChips. Methods Inf Med 2005; 44 (3): 414–417.
18. Choi H, Shen R, Chinnaiyan AM, Ghosh D. A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. BMC Bioinformatics 2007; 8: 364.
19. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003; 19 (2): 185–193.
20. Xu L, Tan AC, Winslow RL, Geman D. Merging microarray data from separate breast cancer studies provides a robust prognostic test. BMC Bioinformatics 2008; 9: 125.
21. Bauer E, Kohavi R. An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. 1999; 36 (1–2): 105–139.
22. Stollhoff R, Sauerbrei W, Schumacher M. An experimental evaluation of boosting methods for classification. Methods Inf Med; 49 (3): 219–229.
23. Yang J, Zhang D, Frangi AF, Yang JY. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans Pattern Anal Mach Intell 2004; 26 (1): 131–137.
24. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143 (1): 29–36.
25. Hailesellasse Sene K, Porter CJ, Palidwor G, Perez-Iratxeta C, Muro EM, Campbell PA, Rudnicki MA, Andrade-Navarro MA. Gene function in early mouse embryonic stem cell differentiation. BMC Genomics 2007; 8: 85.
26. Demsar J. Statistical Comparisons of Classifiers over Multiple Data Sets 2006; 7: 1–30.
27. Mikkelsen TS, Hanna J, Zhang X, Ku M, Wernig M, Schorderet P, Bernstein BE, Jaenisch R, Lander ES, Meissner A. Dissecting direct reprogramming through integrative genomic analysis. Nature 2008; 454 (7200): 49–55.