Gene Expression Deconvolution with Single-cell Data

James Lindsay, Caroline Jakuba, Craig Nelson and Ion Măndoiu

Computer Science and Engineering
Molecular and Cell Biology
University of Connecticut
{james.lindsay,ion}@engr.uconn.edu
{caroline.jakuba@uconn.edu,craig.nelson}@engr.uconn.edu

Abstract. Current in vitro methods for deriving pre-somitic mesoderm have an efficiency of less than 1% indicating an insufficient understanding of mammalian mesodermal differentiation. Here we present an initial look at an approach that augments current digital deconvolution algorithms with single-cell resolution data to derive a better characterization of the transcriptional profiles of cell-types in the embryonic posterior mid-line. This will better enable biologists to direct in vitro derivation methods towards a more developmentally accurate, and therefore more efficient trajectory.

1 Introduction

The pre-somitic mesoderm is the precursor tissue source for all muscle, bone, and cartilage of the mammalian embryo. Given the prevalence of human mesodermal diseases (e.g. muscular dystrophy, osteoporosis, arthritis), generating pre-somitic mesoderm in vitro has great clinical relevance. Unfortunately, current in vitro methods for deriving pre-somitic mesoderm have an efficiency of less than 1% indicating an insufficient understanding of mammalian mesodermal differentiation. Thus far it is known in mouse the pre-somitic mesoderm arises from a self-renewing pool of progenitor cells localized to the posterior midline of the embryo, however, the precise identify of this progenitor cell, i.e. its transcriptional profile, remains unknown.
One approach to identify a cell-types transcriptional profile is in-silico deconvolution of heterogeneous mixtures [5, 9, 6]. Generally the problem is presented as $X_{m\times n} = S_{m\times k}C_{k\times n}$. Where X is a set of gene expression measurements resulting from a linear combination of, S, the cell-type signature matrix and C, the proportions of each cell-type in a particular mixture. There are n mixtures, k cell-types and m genes. Some methods try to estimate both the signature (S) and (C) simultaneously [8, 7, 1]. This *blind* approach to deconvolution is a very difficult task, often only used for separating a small number of cell-types which are very distinct. This approach has little hope to work when there is a complex mixture of similar cell-types as is expected in the mesoderm. Another class of methods assume some information about S is known and try to estimate C [6,

9, 3]. The drawback to this approach is the assumption of that the complete, or partial cell-type profile is known.
The motivation for this paper is to deconvolve population level (i.e. mixtures) qPCR taken from the posterior mid-line of mouse embryos. This was done at the 7-8 somite stage and 12 population qPCR samples were taken, along with 97 single-cells. The expression level of 31 genes was characterized with RT-qPCR. Using the nomenclature described earlier this would give us an expression matrix $X_{31\times12}$. Additionally 97 single-cells from the same embryo section and stage where characterized with RT-qPRC at the same 31 genes. The expression matrix X is given and our goal is to factor this matrix into $S_{31\times k}$ and $C_{k\times12}$. Unfortunately since the true S and C matrix are unknown we have developed a simulation on which to benchmark our approach before applying our approach to the real data.

2 Methods

We propose using a two-pronged approach whereby single-cell qPCR observations are assigned to cell-types using unsupervised clustering in order to create S, then given X and S we can solve for the concentration of each mixture in, C, using constrained quadratic programming.

2.1 Constructing Cell-Type Signatures

Noise Reduction Single-cell qPCR data is known to be noisy due to cell cycle variations, experimental difficulties and due to the fact that the definition of a cell-type is somewhat nebulous [10]. While the third difficulty is quite challenging to address, it is possible to use simple filtering to reduce the noise associated with the first two. In figure 2.2 we used PCA to visualize the variability in the data, we observed little evidence of clustering in (part a). Next we applied a common technique where each sample was required to have .95 Pearson correlation with at-least one sample, otherwise it is removed. There is stronger evidence of clustering (part b).

2.2 Signature Matrix

Clustering gene expression data is a problem that has been widely studied for over a decade. The problem is an instance of unsupervised learning, where samples need to be labelled based on their gene expression. Numerous objectives have been proposed such as minimizing the distance between samples in a cluster, and others focus on grouping functionally related samples. In this developmental context which we are working the proper clustering objective is not immediately evident. The embryo will develop into several distinct regions, and each region will be made up of a mixture of cell-types. Therefore we chose to use k-means clustering to group the single-cell data because it explicitly allows us to control the number of theoretical cell-types. The clustering of the single-cell

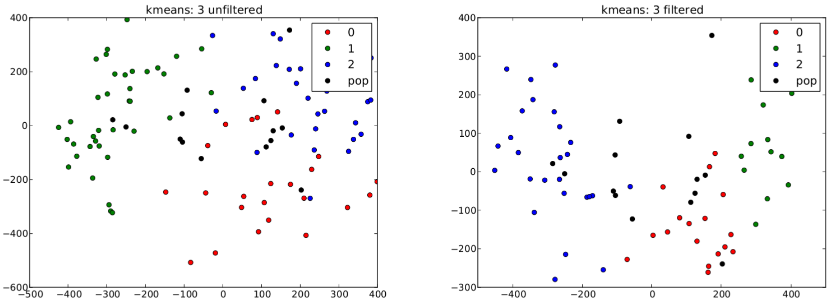

Fig. 1. PCA plot of single-cell qPCR data labelled via k-means cluster with k=3. The population level data is plotted in black. part(a): No filtering was applied. part(b): Pearson correlation filtering is applied.

data can be seen in figure . The average expression profile of each single-cell in a cluster is used to create the cell-type signature matrix $S_{m \times k}$. The cluster assignments have the following proportions, $0 = 29\%$, $1 = 27\%$, $2 = 44\%$.

2.3 Concentration Matrix

The next task is to solve for the concentration matrix $C_{k \times n}$. We utilize the same methodology described in [3] to compute the concentration matrix. Each observation (or column) in the expression matrix X is a linear combination of each cell-type at a given concentration. Let us denote the vector $\bar{x}$ as a particular column in X and the vector $\bar{c}$ as its corresponding column in C.

$$\begin{aligned} \text{minimize} \qquad & ||S\bar{c} - \bar{x}||_2 \\ \text{subject to} \qquad & \sum \bar{c} = 1 \\ \bar{c}_i \geq 0 \ \forall i = 0 \ldots m \end{aligned}$$

This least-squares formulation can be solved with any constrained quadratic programming solver. Once every column in C has been solved the system has been deconvolved.

3 Experimental Results

3.1 Simulation Design

In order to asses our approaches ability to deconvolve qPCR mixtures we developed a way to simulate mixtures. As part of the simulation we will assume that the true signature matrix is known and it is the same as what one would get from utilizing k-means clustering on the single-cell data as described in the previous section, k is fixed to 3. We introduce one parameter into the simulation

$e = 5, 10, 25, 50, 100, 200$ which becomes the noise parameter and it represents the number of single-cell's used to create a particular mixture.
For a single mixture $\hat{x}$ we first simulate the concentration of each cell-type c by choosing k numbers at random from the uniform distribution, and scaling all columns by their sum. Then $e \times \hat{c}_j$ for each cell-type j, gives a count of single-cell observation to choose uniformly at random from each cell-type. $\hat{x}$ is then just a sum of the chosen single-cell. Since our current method is not a *global* optimization, in that each mixture is treated independent of the others we simply chose to use 10 mixtures per experiment yielding an expression matrix $X_{31\times10}$, each experiment was replicated 10 times.

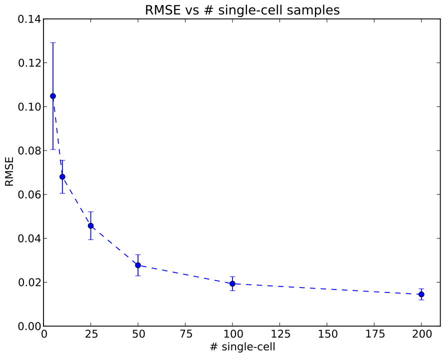

Fig. 2. The average root mean square error (RMSE) for predicted concentrations against the actual is on the y-axis. The x-axis is the number of single-cell observations used in each mixture, which is our *noise* parameter. The error bars represent the standard deviation across the 10 replicates.

3.2 Evaluation Metrics

One common metric for evaluating the accuracy of the constrained least-squares method for estimating concentrations is tracking the root mean square error of each predicted concentration against the truth while varying the number of single-cell mixtures. In figure we see that the RMSE is higher when fewer single-cell samples are taken. However the decrease in RMSE levels off at approximately 100 samples, indicating that more resolution does not help.
While the RMSE is a good summary statistic it is not clear how the cell-type proportions effect the outcome. By plotting the concordance between the observed concentrations and actual for each cell-type we can shed some light. In figure , part a shows that at a high error rate (i.e. few single-cells) there is much more variability between observed and actual concentrations. While in part b, at a low error rate the variability is reduced. However in both parts there is a systematic over and under estimation of concentrations for all cell-types.

3.3 Conclusion

This work is just the first step into fully exploring the use of deconvolution and single-cell resolution qPCR to identify progenitor cell-types in complex mixtures. In figure it was observed that there is a limit to the number of single-cell observations necessary to get meaningful deconvolution. Also figure demonstrates that the least squares approach is capable of deconvolving complex mixtures. However before this method can be applied to real data it is necessary to characterize the effects of skewed cell-type proportions and very rare cell-types like the pre-somitic mesoderm.

Acknowledgements. This project was supported by in part by awards IIS-0546457 and IIS-0916948 from NSF.

References

1. Erkkilä, T., Lehmusvaara, S., Ruusuvuori, P., Visakorpi, T., Shmulevich, I., Lähdesmäki, H.: Probabilistic analysis of gene expression measurements from heterogeneous tissues. Bioinformatics 26(20), 2571–2577 (Oct 2010), http://dx.doi.org/10.1093/bioinformatics/btq406
2. Filho, I.G.C., Filho, C., De Mestrado, D.a., Francisco, O., Carvalho, A.T., Marcílio, C.o., Souto, C.P., De, C., Pós-graduaç ao, I., Ciência, E., Computação, D., Gesteira, I., Filho, C., Data, G.E.: Comparative Analysis of Clustering Methods for Gene Expression Data (2003), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.61.5129
3. Gong, T., Hartmann, N., Kohane, I.S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S., Szustakowski, J.D.: Optimal Deconvolution of Transcriptional Profiling Data Using Quadratic Programming with Application to Complex Clinical Blood Samples. PLoS ONE 6(11), e27156+ (Nov 2011), http://dx.doi.org/10.1371/journal.pone.0027156
4. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: A survey. IEEE Trans. on Knowl. and Data Eng. 16(11), 1370–1386 (Nov 2004), http://dx.doi.org/10.1109/TKDE.2004.68
5. Lu, P., Nakorchevskiy, A., Marcotte, E.M.: Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations. Proceedings of the National Academy of Sciences 100(18), 10370–10375 (Sep 2003), http://dx.doi.org/10.1073/pnas.1832361100
6. Qiao, W., Quon, G., Csaszar, E., Yu, M., Morris, Q., Zandstra, P.W.: PERT: A Method for Expression Deconvolution of Human Blood Samples from Varied Microenvironmental and Developmental Conditions. PLoS Comput Biol 8(12), e1002838+ (Dec 2012), http://dx.doi.org/10.1371/journal.pcbi.1002838
7. Repsilber, D., Kern, S., Telaar, A., Walzl, G., Black, G.F., Selbig, J., Parida, S.K., Kaufmann, S.H., Jacobsen, M.: Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. BMC bioinformatics 11(1), 27+ (Jan 2010), http://dx.doi.org/10.1186/1471-2105-11-27
8. Schwartz, R., Shackney, S.E.: Applying unmixing to gene expression data for tumor phylogeny inference. BMC bioinformatics 11(1), 42+ (Jan 2010), http://dx.doi.org/10.1186/1471-2105-11-42

9. Shen-Orr, S.S., Tibshirani, R., Khatri, P., Bodian, D.L., Staedtler, F., Perry, N.M., Hastie, T., Sarwal, M.M., Davis, M.M., Butte, A.J.: Cell typespecific gene expression differences in complex tissues. Nature Methods 7(4), 287–289 (Mar 2010), http://dx.doi.org/10.1038/nmeth.1439

10. Taniguchi, K., Kajiyama, T., Kambara, H.: Quantitative analysis of gene expression in a single cell by qPCR. Nature methods 6(7), 503–506 (Jul 2009), http://dx.doi.org/10.1038/nmeth.1338

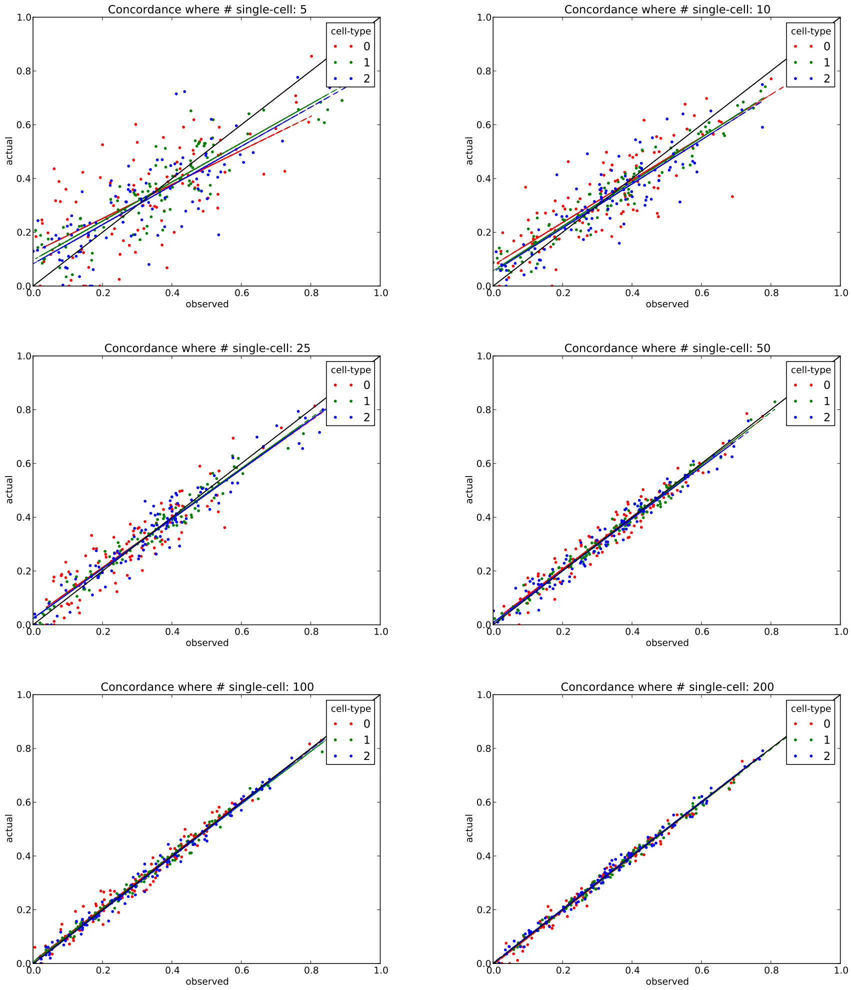

Fig. 3. The concordance between observed and actual concentrations per cell-type. The black line from (0,0) to (1,1) represents perfect concordance. The colored lines is the line of best for its respective cell-type. A different number of single-cells was sampled for each sub-figure.