Statistics and Applications {ISSN 2454-7395 (online)}
Volume 19, No. 1, 2021 (New Series), 431-442

Use of Linear Combination Test to Identify Gene Signatures of Human Embryonic Development in Single Cell RNA-Seq Experiments

I. Dinu[1], E. Khodayari Moez[1], M. Hajihosseini[1], A.P. Leite[2] and S. Pyne[3,4]*

[1]*School of Public Health, University of Alberta, Edmonton, AB, Canada.*
[2]*University College London Cancer Institute, London, UK.*
[3]*Public Health Dynamics Laboratory; and Department of Biostatistics,*
Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA.
[4]*Health Analytics Network, Pittsburgh, PA, USA.*

Received: 05 March 2021; Revised: 06 April 2021; Accepted: 11 April 2021

Abstract

Data on human systems biology are being generated at a rapid pace due to technological advances in not only high-throughput, but also high-resolution, platforms. Increasing availability of single cell omic data have motivated complex experiments with the intention to gain deeper insights into complex biological systems such as those involved in the development of organisms. Individual, technological and biological sources of heterogeneity of outcomes that are observed among the different populations of cells that are sampled in such experiments require robust analysis. We describe our Linear Combination Test (LCT) methodology, and briefly review its applications to binary, multivariate continuous and longitudinal outcomes in a wide range of omic studies. It allows us to test hypothesis not just about the role of single genes in discrete outcomes, but of large sets of genes in multivariate continuous outcomes, which are representative of dynamic biological phenomena such as embryogenesis, degenerative diseases, *etc*. LCT, which uses a shrinkage covariance matrix estimator, has been shown to be effective at a small computational cost in both simulated omic studies and real-life biomedical applications. In this study, we applied LCT to analyze a new collection of stem cell gene signatures associated with single cell RNA-Seq data measured during human preimplantation embryonic development.

Key words: Linear combination test; Single cell analysis, RNA-Seq; Stem cell gene signatures.

1. Introduction

In early microarray data analysis, individual genes that were differentially expressed across 2 or more classes or conditions were identified using traditional statistical methods such as the t-test, ANOVA, *etc*., Drăghici (2012). Then, the most significant genes were selected based on a predefined threshold and validated for biological patterns. However, given the heterogeneity of gene expression levels, biological interpretation of the results was sensitive to the choice of the threshold, and this subjectivity remains an important concern in such analysis of individual genes. In order to overcome this problem, Gene Set Analysis (GSA) uses existing experimentally obtained knowledge of genes and their pathways to test for significant regulation of sets of multiple genes (called *geneset*s) instead of individual genes. Since the genes within such genesets share a common biological function, considering the correlations

Corresponding Author: Saumyadipta Pyne
E-mail: spyne@pitt.edu, spyne@healthanalytics.net

within each set is a key aspect of a sound GSA method. However, it was shown by Tsai and Chen (2009) that many of the known GSA methods, *e.g*., Wang *et al*. (2011), Dinu *et al*. (2007), Goeman *et al*. (2004), Mansmann and Meister (2005), Kong *et al*. (2006), Subramanian (2005), Efron and Tibshirani (2007), were affected by large type II errors. Another important limitation of many GSA methods is that they can only accommodate binary outcomes, such as disease versus control. Our method, Linear Combination Test (LCT) is a GSA method that was designed to address these limitations by taking into account correlations across genes and outcomes, and dealing with binary, univariate or multivariate continuous outcomes, measured either at a single point in time or at multiple time points, and therefore, allow us to analyze a wider range of studies involving complex study designs (Wang *et al*. 2014).

Single cell omic studies have become increasingly popular over the past decade, due to their powerful ability to profile from a panel of selected few dozen genes up to the entire transcriptome of a large number of individual cells in parallel. A typical example of a basic experiment on single-cell analysis (SCA) of gene expression is described in Figure 1 due to Kalisky *et al*. (2018). SCA involves experiments on individual cells that are typically isolated from a biological sample and then physically separated from each other and assayed upon DNA amplification. For each cell, the expressions of genes (or their products such as proteins) are measured using different well-established (or newly emerging) omic platforms such as RNA sequencing as reviewed by Lafzi (2018), Dal Molin (2019), Kalisky (2018). In the resulting data, the single cells could be considered as samples that are assumed independent and hence possibly affected by different sources of variation among the expression-levels of even the same genes.

The large volume of data measured by single cell omic studies calls for sound statistical and computational analysis methods. Various methods at the individual gene level, have been reviewed by Andrews *et al*. (2021). While most of such methods are focused on differential expression of individual genes between cells representing (generally two) different states, here we reason that an analysis using sets of genes, *i.e*., GSA, has important advantages over the individual gene level analysis when applied to single cell omic studies. The stochasticity of expression levels of the same genes in individual cells could be due to different factors ranging from biological (*e.g*., the cell cycle phase of a particular cell) to technical (*e.g*., missing data). While specific genes may not show consistent expression across single cells, if we consider testing the differential expression of not one gene at a time but multiple genes together that are known to belong to a carefully selected geneset, then our LCT method is more likely than traditional approaches to detect the regulation of a functional process or biological pathway that is significantly associated with the outcomes of a given SCA experiment.

Interestingly, LCT allows multivariate and continuous outcomes that could be more realistic representations of single cell level stochasticity of behaviors than univariate and discrete class labels as used in traditional studies of bulk samples. For such reasons, LCT can provide an overall more robust analytical approach for SCA experiments. In addition, LCT type I error, power and computational efficiency were compared to top GSA methods in simulations and real data analysis studies (Wang *et al*., 2014). LCT type I error and power were comparable to MANOVA-GSA (Tsai and Chen 2009), and superior to SAM-GS (Dinu *et al*., 2007), especially at higher magnitudes of the correlations values across sets of genes, which is a common scenario in GSA. LCT was superior to both methods in terms of computational efficiency. LCT performed better than GSEA in a simulation study presented by Khodayari *et al*. (2018). However, we would like to point out that GSEA uses information from genes exterior to a pre-defined set or pathway. Based on methodological considerations, Goeman and

Buhlmann (2007) discourage comparing methods involving only genes belonging to the pre-defined set to methods involving genes outside the pre-defined set, as these two categories are conceptually different, and they are testing different hypotheses.

GSA focuses on analysis of biological pathways, or genesets sharing a common biological function. Well-known examples of such collections of pre-defined, often expert-curated, genesets include The Cancer Genome Atlas (TCGA), Tomczak *et al.* (2015); Gene Expression Omnibus (GEO), Edgar *et al.* (2002); Kyoto Encyclopedia of Genes and Genomes (KEGG), Kanehisa *et al.* (2000); BioCarta, Nishimura *et al.* (2001); Molecular Signature Database (MSigDB), Liberzon *et al.* (2015). The use of a carefully selected collection of genesets relevant to the outcomes of interest is a key aspect in GSA. In this study, we compiled a new, large collection of genesets that were reported by several past embryonic stem cell gene expression experiments, and used them to test their association with different stages of early human embryonic development. The remaining of this paper consists of a presentation of the LCT methodology, followed by its application to single cell embryonic genome-wide expression (RNA-Seq) data. We will also discuss various extensions of LCT, including its applications beyond gene expression studies.

2. Data and Methods

Data: In this study, we used 2 types of data. First, for our genesets, we introduce a large collection of 457 curated genesets that were derived from experimentally identified signatures of gene expression in human embryonic stem cells. Hence, we call the collection “stem cell signatures”. These genesets were compiled from the Molecular Signature DataBase (MigDB), Liberzon *et al.* (2015); the Differentiation Map portal, Novershtern *et al.* (2011); Ingenuity Pathway Analysis tool (2020); and ChIP-X database, Lachmann *et al.* (2010). We restricted the size of genesets to be between 5 and 500. There are 281 genesets in this range (The full collection of these stem cell signatures is available from the authors upon request). Second, we downloaded the single cell RNA-Seq data from ArrayExpress database, Athar *et al.* (2019); access number E-MTAB-3929 ArrayExpress (2020). The dataset consists of 17855 genes measured in each of 1529 individual cells from 88 human preimplantation embryos. The total of 1529 individual cells is broken down during days 3 to 7 of the embryonic development as follows: 81 cells measured on day 3; 190 on day 4; 377 on day 5; 415 on day 6; and 466 on day 7. During the first 7 days of human development, the zygote undergoes cellular division and establishes the first three distinct cell types of the mature blastocyst: trophectoderm (TE), primitive endoderm (PE), and epiblast (EPI). Petropoulos *et al.* (2016) While the analysis of these data at individual gene level was conducted previously by Petropoulos (2016), in this study, we performed LCT analysis at geneset level of the same data using the above-mentioned stem cell signatures.

Methods: LCT tests if there is a linear relationship between the geneset $X = \{x_1, \ldots, x_p\}$ consisting of p genes and a set of q multivariate outcomes $Y = \{Y_1, \ldots Y_q\}$. The multivariate null hypothesis can be expressed linearly and univariately as

H_0: There is no association between any linear combination of gene expressions of the members of a geneset X and any linear combination of multivariate outcomes Y.

If $Z(X, A)$ is a linear combination of gene expression measurements within a set of x_is with coefficient vector A and $Z(Y, B)$ is a linear combination of outcomes y_is with coefficient vector B, then we calculate the following statistic to test the null hypothesis

$$T^2 = max|\rho((Z(X,A),Z(Y,B))^2|. \quad (1)$$

The coefficient vectors A and B are estimated in a way that maximizes the Pearson correlation between $Z(X,A)$ and $Z(Y,B)$. T^2 can be rewritten as

$$T^2 = max\frac{\left(A^T Cov(X,Y)B\right)^2}{(A^T Cov(X,X)A).(B^T Cov(Y,Y)B)} = \frac{\left(A^T \Sigma_{XY} B\right)^2}{(A^T \Sigma_{XX} A)\,.(B^T \Sigma_{YY} B)}\,. \quad (2)$$

In the procedure for estimation of the coefficient vectors, two problems arise: singularity caused by the high dimensionality of data (solved by shrinkage methods) and computational efficiency (solved by eigenvalue decomposition). Then, the p-value is calculated using sample permutations. Sample permutation method preserves the correlation structure within geneset and the correlation structure within multivariate outcomes, see Schäfer and Strimmer (2005).

Specifically, the (i,j)th entry of the shrinkage covariance matrix Σ_{XX}^* is given by

$$\sigma_{ij}^* = \gamma_{ij}\sqrt{\sigma_{ii}\sigma_{jj}}$$

with shrinkage coefficients 1 for the diagonal terms, and the off-diagonal terms

$$\gamma_{ij} = \rho_{ij}\, min(\,1, max(\,0{,}1 - \lambda^*))$$

where ρ_{ij} is the sample correlation between x_i and x_j. The optimal shrinkage intensity can be estimated by

$$\lambda^* = \frac{\sum_{i\neq j} var(\rho_{ij})}{\sum_{i\neq j} \rho_{ij}^2}\,.$$

Based on this shrinkage strategy, we get the shrinkage version of the test statistic

$$T^{2*} = \max_{A,B}\frac{(A^T\Sigma_{XY}B)^2}{(A^T\Sigma_{XX}^*A)\,.\,(B^T\Sigma_{YY}^*B)}. \quad (3)$$

The computational cost of calculating (3) has to be taken into consideration, since the right-hand side is a nonlinear programming problem involving $p+q$ parameters. The computational cost can be very high for maximizing directly the right-hand side of (3), especially when permutations are used for calculating the p-values of the test. To address the computational efficiency problem, we adopt a strategy of using two groups of normalized orthogonal bases, instead of using the original observation vectors of X and Y. We perform eigenvalue decompositions for the two shrinkage covariance matrices, $\Sigma_{XX}^* = UD_XU^T$ and $\Sigma_{YY}^* = VD_YV^T$, and obtain two groups of orthogonal basis vectors $\tilde{X} = (\tilde{x}_1 \ldots, \tilde{x}_p) = (x_1 - \bar{x}_1, \ldots, x_p - \bar{x}_p)UD_X^{-1/2}$ and vectors $\tilde{Y} = (\tilde{y}_1 \ldots, \tilde{y}_q) = (y_1 - \bar{y}_1, \ldots, y_q - \bar{y}_q)VD_Y^{-1/2}$.

The test statistic in (3) can further be rewritten as

$$T^{2*} = \max_{\alpha,\beta}\frac{(\alpha^T\Sigma_{\tilde{X}\tilde{Y}}\beta)^2}{||\alpha||_2^2 \cdot ||\beta||_2^2}\,, \quad (4)$$

where $\alpha = D_X^{1/2} U^T A$ and $\beta = D_Y^{1/2} V^T B$, and $\Sigma_{\tilde{X}\tilde{Y}}$ is the covariance matrix between $\tilde{X}$ and $\tilde{Y}$, with its (i, j)th entry being $cov(\tilde{x}_i, \tilde{y}_j)$.

The optimization problem in (4) can be solved in two steps. First, for a given β, we find the optimal, α which is proportional to $\Sigma_{\tilde{X}\tilde{Y}}\beta$; second, substitute the optimal α into (4), and find the global optimal, which is proportional to the first eigenvector of the matrix $\Sigma_{\tilde{X}\tilde{Y}}^T \Sigma_{\tilde{X}\tilde{Y}}$ corresponding to the largest eigenvalue. We note that the value of T^{2^*} equals to the largest eigenvalue of either the $q \times q$ matrix $\Sigma_{\tilde{X}\tilde{Y}}^T \Sigma_{\tilde{X}\tilde{Y}}$ or the $p \times p$ matrix $\Sigma_{\tilde{Y}\tilde{X}}^T \Sigma_{\tilde{Y}\tilde{X}}$. The cost of obtaining the largest eigenvalue is low, providing $\min(p, q)$ is not large.

The computational advantage is obvious when sample permutations are used to calculate p-values of the test. Since sample permutation changes neither the correlation structure within genesets nor that within the outcomes, we do not need to repeat the same eigenvalue decompositions of the two shrinkage covariance matrices in (3) for the permuted data, but only for the original outcome. That is, after performing the eigenvalue decompositions for the two shrinkage covariance matrices Σ_{XX}^* and Σ_{YY}^* and creating two groups of orthogonal basis vectors $\tilde{X}$ and $\tilde{Y}$, permutations can be done only on $\tilde{Y}$ directly, instead of on the original outcome Y.

For multiple comparisons over large collections of genesets, False Discovery Rate (FDR) is a commonly used method that can provide a better alternative to the more conservative Bonferroni approach. In this study, we used q-value, which is the expected positive FDR, to identify the significantly regulated genesets at different q-value levels (Storey and Tibshirani 2003).

3. Results and Discussion

In this study, we used LCT for testing associations of a new, large collection of curated stem cell signatures with a single-cell RNA-Seq based genome-wide expression dataset on human embryo development. We conducted a quick confirmation of the relevance of these signatures in stem cell gene regulation during human embryo development by applying LCT to single cell data across each pair of consecutive days, from day 3 to day 7. Petropoulos et al. (2016) reported results of an analysis at the individual gene level, across the three distinct cell types of the mature blastocyst. We note that segregation of EPI, PE and TE cell types appears at day 5. The breakdown of sample sizes by day and cell type is as follows: 41 EPI, 32 PE and 142 TE for day 5; 45 EPI, 39 PE and 331 TE for day 6; 41 EPI, 37 PE and 388 TE for day 7.

Differential expression analysis between the EPI cell types and PE cell types performed by Petropoulos *et al.* (2016) identified 43, 1,412, and 542 differentially expressed genes at days 5, 6 and 7 respectively (at FDR ≤ 0.05), with earlier days' (5 and 6) significance being maintained through later days (6 and 7). Our analysis at the geneset level identified 126 differentially regulated stem cell signatures between EPI and PE at day 5 (q-value ≤ 0.001), and a selected subset of 105 genesets that were the most significant (q-value ≤ 0.0001) on days 6 and 7 is shown in Table 1. Regarding the other two cell type pairs, TE versus PE, and TE versus EPI, our analysis at the geneset level indicated more obvious differences compared to EPI versus PE. More importantly, all 281 genesets, which are known stem cell signatures, were found to be significant on each pair of consecutive days from 5 to 7, and for each pair of cell types of TE versus PE, and TE versus EPI, which is in agreement with the individual gene analysis results of Petropoulos et al. (2016).

Finally, we focused on the endothelial transcription factor genes *GATA2* and *GATA3*, which have been previously reported as known markers of TE segregation, and thus, for playing an important role in embryonic development. Ortega *et al.* (2018) We performed LCT using the expressions of these two genes as a bivariate continuous outcome. One hundred and thirty-eight stem cell signatures were significantly associated (q-value $\leq$ 0.001) with the bivariate continuous phenotype at day 3. For the subsequent days, this count increased to 234, 278 and 281 respectively.

We presented here a new application of LCT methodology to SCA experiments with an illustration on human embryonic development gene expression data. Our approach extends the individual gene analysis to identification of sets of genes that share a common biological function. Such collections of sets exhibit higher reproducibility across studies, and are more robust for addressing complex questions in systems biology. Notably, our new collection of stem cell signatures presented the opportunity to confirm their relevance to the dynamic gene regulation during human embryonic development. As stem cells are an active area of research in biology and medicine, multivariate dynamic outcomes and associated markers (and combinations thereof) can be analyzed by LCT, which can also be extended to testing of newer gene signatures such as those reported by novel experiments to chart a single cell level transcriptional roadmap of human development, *e.g.*, Blakeley *et al.* (2015), Durruthy-Durruthy *et al.* (2016).

Unlike the traditional analysis of bulk samples composed of thousands of different cells, experiments that can measure the expressions of selected markers in individual cells are capable of revealing not only the occurrence but also the dynamic states of diverse cell populations, including rare ones, as shown by Pyne *et al.* (2009), Pyne *et al.* (2014), Qi *et al.* (2020), *etc*. In order to characterize the cellular heterogeneity of a given sample (say, a tumor) with precision, the experimenter will need to select the corresponding panels of marker genes. Different choices of markers are given by genesets which must be compiled carefully as we have done for characterizing embryonic stem cell signatures in this study.

LCT is a powerful correlation-based test that can be used to explore thousands of genesets in an automated yet computationally efficient manner. However, we note that the linear combinations identified by LCT are not unique, and there is no direct interpretation of the linear combination coefficients that one can achieve through classical linear regression techniques. To address this, statistical models for high-dimensional analysis can be applied *post hoc* to the identified sets of genes, based on their LCT significance, and selected for their biological relevance. Such methods may provide further interpretation and insight into the selected sets. LCT was extended to longitudinal multivariate outcomes by Khodayari *et al.* (2019). LCT has also proven to be effective beyond gene expression data. Analogous to genesets, we have used LCT on collections of metabolite-sets to test for associations between oncogenic outcomes and high-throughput metabolomic data from prostate cancer patients in Khodayari *et al.* (2018).

References

Andrews, T. S., Kiselev, V. Y., McCarthy, D. and Hemberg, M. (2021). Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nature Protocols*, **16**, 1-9.

ArrayExpress. (2020). https://www.ebi.ac.uk/arrayexpress/.

Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N. A., Petryszak, R., Paptheodorou, I., Sarkans, U. and Brazma, A. (2019). ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Research*, **47**(**D1**), D711-D715.

Blakeley, P., Fogarty, N. M., del Valle, I., Wamaitha, S. E., Hu, T. X., Elder, K., Snell, P., Christie, L., Robson, P. and Niakan, K. K. (2015). Defining the three cell lineages of the human blastocyst by single-cell RNA-Seq. *Development*, **142**(**18**), 3151-3165.

Dal Molin, A. and Di Camillo, B. How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. (2019). *Brief Bioinformatics*, **20**(**4**),1384-1394.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri G. S., Einecki, G., Famulski, K. S., Halloran, P., Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, **8**, 242

Drăghici, S. (2012). *Statistics and Data Analysis for Microarrays Using R and Bioconductor. 2nd Edition*. Chapman & Hall/CRC Mathematical and Computational Biology Series (ISSN: 978-1439809754).

Durruthy-Durruthy, J., Wossidlo, M., Pai, S., Takahashi, Y., Kang, G., Omberg, L., Chen, B., Nakauchi, H., Reijo Pera, R. and Sebastiano, V. (2016). Spatiotemporal reconstruction of the human blastocyst by single-cell gene-expression analysis informs induction of Naive Pluripotency. *Development Cell*, **38**(**1**), 100-115.

Edgar, R., Domrachev, M. and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, **30**, 207–210.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *Annals of Applied* Statistics, **1**, 107–129.

Goeman, J. J. and Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980-987.

Goeman J. J., van de Geer S. A., de Kort, F. and van Houwelingen, H. C. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20,** 93–99.

Ingenuity Pathway Analysis Tool. (2020). http://www.ingenuity.com.

Kalisky, T., Oriel, S., Bar-Lev, T. H., Ben-Haim, N., Trink, A., Wineberg, Y., Kanter, I., Gilad, S. and Pyne, S. (2018). A brief review of single-cell transcriptomic technologies. *Brief Functional Genomics*, **17**(**1**), 64-76.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **28**, 27–30.

Khodayari, M. E., Hajihosseini, M., Andrews, J. L. and Dinu, I. (2019). Longitudinal linear combination test for gene set analysis. *BMC Bioinformatics*, **20**(**1**), 650.

Khodayari, M. E., Pyne, S. and Dinu, I. (2018). Association between bivariate expression of key oncogenes and metabolic phenotypes of patients with prostate cancer. *Computers in Biology and Medicine*, **103**, 55-63.

Kong, S. W., Pu, W. T. and Park, P. J. (2006). A multivariate approach for integrating genome wide expression data and biological knowledge. *Bioinformatics*, **22,** 2373-2380.

Lachmann, A., Xu, H., Krishnan, J., Berger, S. I., Mazloom, A. R. and Ma'ayan, A. (2010). ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**(**19**), 2438-2444.

Lafzi, A., Moutinho, C., Picelli, S. and Heyn, H. (2018). Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Natural Protocols*, **13**(**12**), 2742-2757.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P. and Tamayo, P. (2015). The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Systems*, **1**(**6**), 417-425.

Mansmann, U. and Meister, R. (2005). Testing differential gene expression in functional groups. *Methods of Information in Medicine*, **44**, 449-53.

Nishimura, D. (2001). BioCarta. *Biotech Software and Internet Report*, **2**, 117–120.

Novershtern, N., Subramanian, A., Lawton, L. N., Mak, R. H., Haining, W. N., McConkey, M. E., Habib, N., Yosef, N., Chang, C. Y., Shay, T., Frampton, G. M., Drake, A. C., Leskov, I., Nilsson, B., Preffer, F., Dombkowski, D., Evans, J. W., Liefeld, T., Smutko, J. S., Chen, J., Friedman, N., Young, R. A., Golub, T. R., Regev, A. and Ebert, B. L. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, **144**(**2**), 296-309.

Ortega, M. N., Winblad, N., Reyes, A. P., Lanner, F. (2018). Functional genetics of early human development. *Current Opinion in Genetics and Development*, **52**, 1-6.

Petropoulos, S., Edsgard, D.,, Reinius, B., Linnarsson, S., Sandberg, R. and Lanner, F. (2016). Single-cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell*, **165**, 1012-1026.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T-I., Maier, L. M., Beacher-A. C., *et al*. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences USA*, **106**(**21**), 8519-8524.

Pyne, S., Lee, S. X., Wang, K., Irish, J., Tamayo, P., Nazaire, M-D., Duong, T., Ng, S-K., Hafler, D., *et al*. (2014). Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. *PLoS ONE*, **9**(**7**), e100334.

Qi, Y., Fang, Y., Sinclair, D., Guo, S., Alberich-Jorda, M., Lu, J., Tenen, D., Kharas, M. and Pyne, S. (2020). High-speed automatic characterization of rare events in flow cytometric data. *PLoS ONE*, **15**(**2**): e0228651.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**, 32.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy S. L., Golub T. R., Lander, E. S. and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**, 15545–15550.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440–9445.

Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015). The cancer genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*, **19**, A68–A77.

Tsai, C. and Chen, J. J. (2009). Multivariate analysis of variance test for geneset analysis. *Bioinformatics.* **25**(**7**), 897-903.

Wang, X., Dinu, I., Liu, W. and Yasui, Y. (2011). Linear combination test for Hierarchical Gene Set Analysis. *Statistical Applications in Genetics and Molecular Biology*, **10**(**1**), Article 13.

Wang, X., Pyne, S. and Dinu, I. (2014). Gene set enrichment analysis for multiple continuous phenotypes. *BMC Bioinformatics*, **15**, 260.

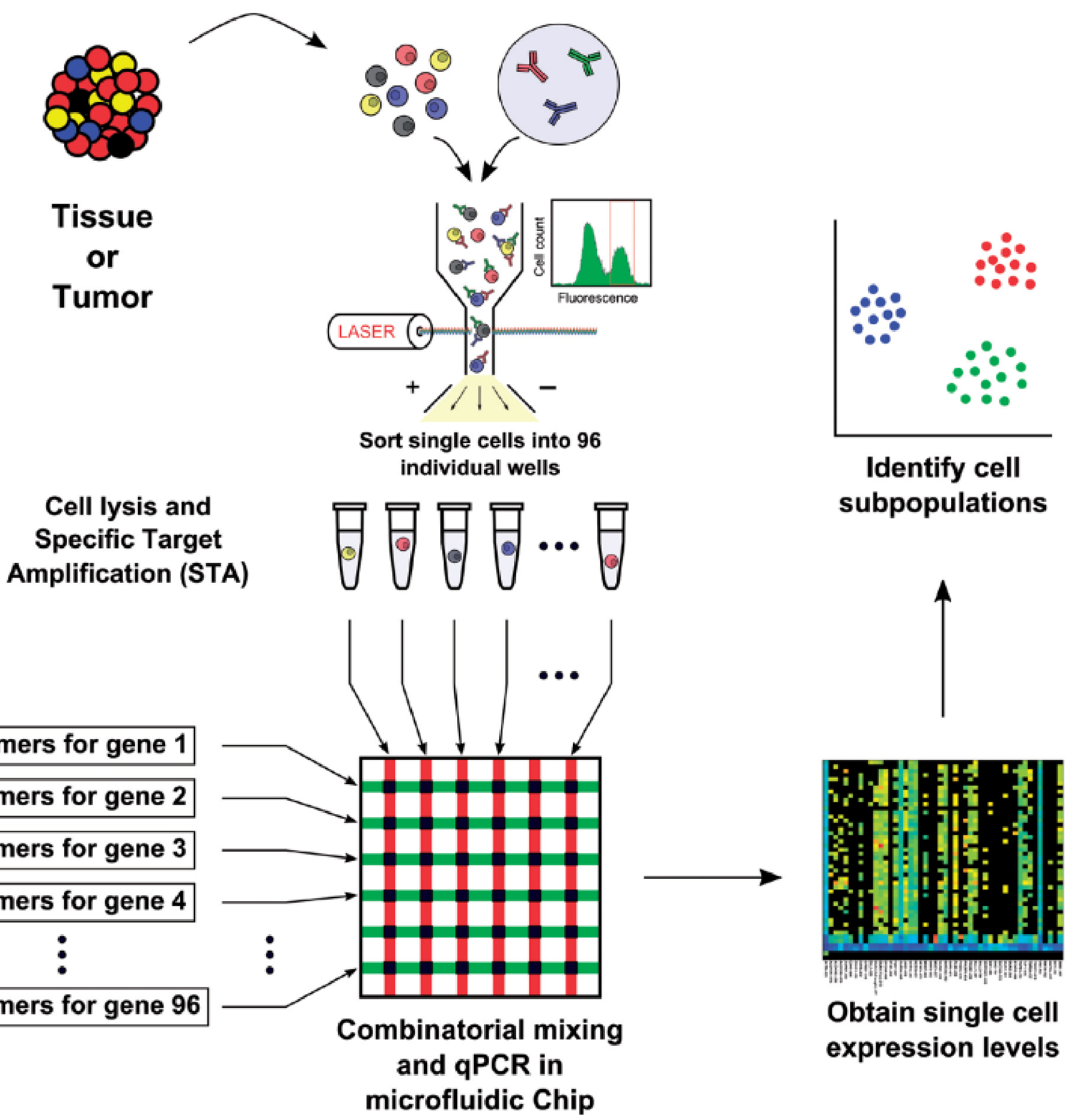
Tissue
or
Tumor
Cell count
Fluorescence
LASER
+
−
Sort single cells into 96 individual wells
Identify cell subpopulations
Cell lysis and Specific Target Amplification (STA)
Primers for gene 1
Primers for gene 2
Primers for gene 3
Primers for gene 4
Primers for gene 96
Combinatorial mixing and qPCR in microfluidic Chip
Obtain single cell expression levels

Figure 1: A schematic diagram showing a basic SCA experiment. Single-cell gene expression measurement using qPCR workflow is performed with the Fluidigm Dynamic Array microfluidic chip. Reproduced from Kalisky *et al.* (2018) with permission.

Table 1: Stem cell gene signatures differentially regulated between 41 epiblast and 32 primitive endoderm lineages at day 5 in human embryo development (shown in the increasing order of size)

Geneset	Size	*q*-value
StemCell_Lian07_20genes_17053208.table1a	20	<0.0001
StemCell_Shim04_23genes_15246160.table4	22	<0.0001
IPA_affects.epithelial.mesenchymal.transition.of.cells	22	<0.0001
IPA_increases.epithelial.mesenchymal.transition.of.cells	24	<0.0001
StemCell_Kocer08_44genes_18667080.TableS3	26	<0.0001
StemCell_Matushansky08_35genes_18310505.TableS6	28	<0.0001
StemCell_Lottaz10_30genes_20145155.Table1	29	<0.0001
IPA_increases.differentiation.of.embryonic.stem.cells	29	<0.0001
StemCell_Lim08_35genes_18510698.Table3	34	<0.0001
Ben.Porath_ES_2	39	<0.0001
StemCell_Seo07_61genes_18034892.Table1	40	<0.0001
DMAP_TCELLA3_DN	40	<0.0001
Marson_H3K4me3	41	<0.0001
DMAP_PRE_BCELL2_UP	42	<0.0001
DMAP_PRE_BCELL3_DN	42	<0.0001
IPA_affects.differentiation.of.embryonic.stem.cells	43	<0.0001
DMAP_EOS_DN	44	<0.0001
DMAP_BCELLA3_DN	44	<0.0001
DMAP_CMP_DN	44	<0.0001
DMAP_NKA2_DN	44	<0.0001
DMAP_TCELL_DN	45	<0.0001
DMAP_GRAN3_UP	45	<0.0001
DMAP_MEGA2_DN	45	<0.0001
DMAP_NKA3_UP	45	<0.0001
DMAP_ERY1_DN	46	<0.0001
DMAP_ERY2_UP	46	<0.0001
DMAP_HSC3_DN	46	<0.0001
DMAP_BCELLA1_DN	47	<0.0001
DMAP_TCELLA2_UP	47	<0.0001
DMAP_TCELLA6_DN	47	<0.0001
StemCell_Lim08_50genes_18510698.Table1	48	<0.0001
DMAP_MYP_UP	48	<0.0001
DMAP_BCELLA2_UP	49	<0.0001
StemCell_Duhagon10_60genes_20500816.Table1	56	<0.0001
IPA_affects.differentiation.of.hematopoietic.progenitor.cells	56	<0.0001
IPA_affects.differentiation.of.hematopoietic.cells	62	<0.0001
IPA_increases.differentiation.of.stem.cells	68	<0.0001
StemCell_Kocer08_87genes_18667080.TableS6	71	<0.0001
IPA_affects.differentiation.of.stem.cells	73	<0.0001

StemCell_Hao09_97genes_20077526.TableS7	74	<0.0001
Kim_CORE	74	<0.0001
StemCell_Almstrup04_138genes_15256440.tableS1	80	<0.0001
DB_KLF2.18264089	83	<0.0001
DB_KLF4.18264089	83	<0.0001
DB_KLF5.18264089	83	<0.0001
Taube.et.al._EMT_upregulated_genes	86	<0.0001
StemCell_Colombo09_111genes_19123479.TableS1	90	<0.0001
IPA_affects.differentiation.of.bone.marrow.cells	90	<0.0001
DB_TRP63.18441228	94	<0.0001
Ben.Porath_ES_CORE_NINE_CORRELATED	99	<0.0001
StemCell_Korkola05_146genes_15870693.SuppTable1	100	<0.0001
StemCell_Bohgaki05_118genes_16014681.table2	113	<0.0001
DB_NOTCH1.17114293	121	<0.0001
StemCell_Hao09_173genes_20077526.TableS5	126	<0.0001
DB_HOXD13.18407260	130	<0.0001
StemCell_Kocer08_185genes_18667080.TableS4	154	<0.0001
Ben.Porath_NOS_TARGETS	168	<0.0001
DB_IRF1.19129219	173	<0.0001
DB_TP63.19390658	176	<0.0001
DB_ESR1.20079471	187	<0.0001
DB_PPARG.19300518	187	<0.0001
StemCell_Kocer08_236genes_18667080.TableS8	194	<0.0001
DB_VDR.20736230	196	<0.0001
DB_WT1.19549856	197	<0.0001
DB_SCL.19346495	206	<0.0001
Kim_GCN5L2	211	<0.0001
Ben.Porath_MYC_TARGETS_WITH_EBOX	222	<0.0001
StemCell_Hao09_359genes_20077526.TableS4	239	<0.0001
DB_EGR1.19032775	242	<0.0001
DB_CDX2.19796622	253	<0.0001
Kim_CTR9	258	<0.0001
StemCell_Matushansky08_297genes_18310505.TableS8	261	<0.0001
DB_ZIC3.20872845	266	<0.0001
Ben.Porath_OCT4_TARGETS	272	<0.0001
StemCell_Hao09_612genes_20077526.TableS3	302	<0.0001
StemCell_Matushansky08_886genes_18310505.TableS1	306	<0.0001
StemCell_Bhattacharya05_2471genes_16207381.Table1Sb	308	<0.0001
StemCell_Kocer08_575genes_18667080.TableS9	309	<0.0001
StemCell_Kocer08_864genes_18667080.TableS2	310	<0.0001
StemCell_Bhattacharya05_2843genes_16207381.Table1Sa	310	<0.0001
StemCell_Majeti09_3024genes_19218430.TableS3	313	<0.0001
StemCell_Matushansky08_1453genes_18310505.TableS7	315	<0.0001
DB_POU5F1.18700969	322	<0.0001

DB_RARG.19884340	326	<0.0001
Zola07_426genes_CellDifferentiationMarkers_17174972.TableS1	332	<0.0001
StemCell_Qi03_534genes_12631704.table1	334	<0.0001
StemCell_Hassan09_1544genes_19808871.TableS3	340	<0.0001
DB_ESR1.17901129	347	<0.0001
DB_TP53.16413492	349	<0.0001
DB_HTT.18923047	355	<0.0001
Ben.Porath_ES_1	358	<0.0001
DB_SMAD4.19686287	367	<0.0001
DB_STAT6.20620947	369	<0.0001
DB_SOX2.18555785	388	<0.0001
DB_CLOCK.20551151	399	<0.0001
DB_NANOG.18555785	412	<0.0001
DB_CTNNB1.20615089	416	<0.0001
DB_TCF4.18268006	420	<0.0001
DB_POU5F1.18555785	438	<0.0001
DB_ZFP281.18358816	441	<0.0001
Kim_PRC	444	<0.0001
DB_CDX2.20551321	446	<0.0001
Kim_ZFP281	461	<0.0001
DB_SMAD1.18555785	465	<0.0001
DB_PDX1.19855005	493	<0.0001