Vol.2012-BIO-29 No.18
2012/6/29

Analysis of single cell gene expression of mESC and MEF by NGS

Kohei Iijima[1] Y-h. Taguchi[1,a),b)]

Abstract: Gene expression development during cell differentiation is a key factor to understand the mechanism of development. However, conventional gene expression analysis cannot distinguish among individual cell expression. In this paper, we re-analyze single cell gene expression measurements obtained by next gene sequencing technology during differentiation from mouse ES cell to MEF.

1. Introduction

Gene expression regulation during cell differentiation is a key factor to understand development[1]. However, gene expression diversity among individual cells is not fully understood. Recently, single cell gene expression technology was invented[2]. Using this technology, one can see how gene expression in individual cells can change during differentiation. Islam *et al*[3] has investigated mice cell gene expression during differentiation from embryonic stem cell (mESC) to mouse embryonic fibroblast (MEF) using next generation gene expression (NGS) technology. We reanalyze their data and try to find individuality of gene expression of each cell.

2. Materials and Methods

2.1 Gene expression data

Gene expression data (GEO : GSE29087) was downloaded as fastq files[*1]. It contains 48 mESC cell samples and 44 MEF samples. Each sample typically consists of three to four fastq files.

2.2 Mapping short read to genome

Because of shortness of read length (58), popular analytical tools like TopHat[4] and Cufflinks[5] could not work very well. In stead of them, we have employed MapSplice[6] and generated SAM files were treated by SAMMate[7]. RPKM attributed to gene symbol and/or refseq mRNA gene id were used for further analysis.

2.3 Gene expression atlas

Gene expression atlas [8] was used in order to check if picked up genes are differently expressed between mESC and MEF.

2.4 SAM

SAM[9] was employed to select genes which are expressive differently between mESC and MEF. SAM was performed by sam function in siggenes package[10] in R[11].

2.5 Principal component analysis

Principal component analysis (PCA) was applied to obtained gene expression profiles. Genes which were embedded far from origin in PCA embeddings were picked up in the first four dimensional embeddings for each sample which was also aligned far from origin in PCA embeddings. PCA was performed by prcomp or princomp function in base package of R[11].

3. Results

In order to select genes which express the difference between mESC and MEF, we employed SAM. Table 1 shows 10 genes with smaller *P*-values corrected by BH criterion. Although only two have significant *P*-values, gene expression atlas gives us concurrent results; most genes are differently expressed between embryo and adult cells (Fig. 1). The reason why we could not get significant *P*-values was because genes were not detected for sufficiently large number of cells.

Since SAM could not give us significant *P*-values, we consult other feature selection methods which are free from *P*-values: i.e., principal component analysis (PCA). Fig. 2 shows PCA embeddings of samples and genes. Although mESCs are clearly placed far from origins, neither mESC nor MEF is clustered (Figs. 2 a and b). In order to see which gene expression push mESC outwards, we have selected outer-most genes in PCA embeddings of genes (Figs. 2 c and d). In Fig. 3, we showed gene expression extracted from gene expression atlas for genes placed far from origin in Fig. 2 c and d. Selected genes are substantially different from genes by SAM and mostly expressive in mESC and suppressive in MEF. Thus, PCA based gene selection turned out to be able to work as a supplementary method for SAM based gene selection.

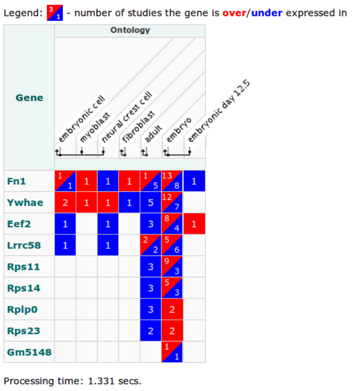
Some genes selected by SAM or PCA are mostly important.

[1] Department of Physics, Chuo University, Tokyo 112–8551, Japan
[a)] tag@granular.com
[b)] Corresponding Author

[*1] http://trace.ddbj.nig.ac.jp/DRASearch/study?acc=SRP006834

Table 1 Frequency of detection of genes with smaller q-values (FDR adjusted P-values).

Gene symbol	mESC	MEF	q-values
Fn1	7	9	0.005
Lrrc58	3	6	0.005
Gm5148	16	14	0.099
Rplp0	9	10	0.099
Rps23	10	8	0.099
Gm15450	10	7	0.099
Rps11	7	8	0.099
Ywhae	6	7	0.099
Eef2	5	8	0.099
Rps14	6	6	0.099



Legend: 🟥1🟦1 - number of studies the gene is over/under expressed in

Gene	embryonic cell	myoblast	neural crest cell	fibroblast	adult	embryo	embryonic day 12.5
Fn1	🟥1 🟦1	🟥1	🟦1	🟥1	🟥1 🟦5	🟥13 🟦8	🟦1
Ywhae	🟥2	🟥1	🟥1	🟦1	🟦5	🟥12 🟦7	
Eef2	🟦1		🟦1		🟦3	🟥8 🟦4	🟥1
Lrrc58	🟦1		🟦1		🟥2 🟦2	🟥5 🟦6	
Rps11					🟦3	🟥9 🟦3	
Rps14					🟦3	🟥5 🟦3	
Rplp0					🟦3	🟥2	
Rps23					🟦2	🟥2	
Gm5148						🟥1 🟦1	

Processing time: 1.331 secs.

Fig. 1 Gene expression profile in gene expression atlas for genes selected by SAM.

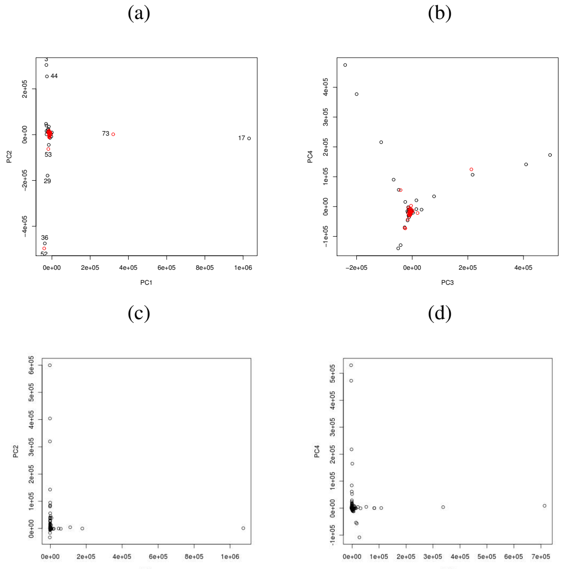
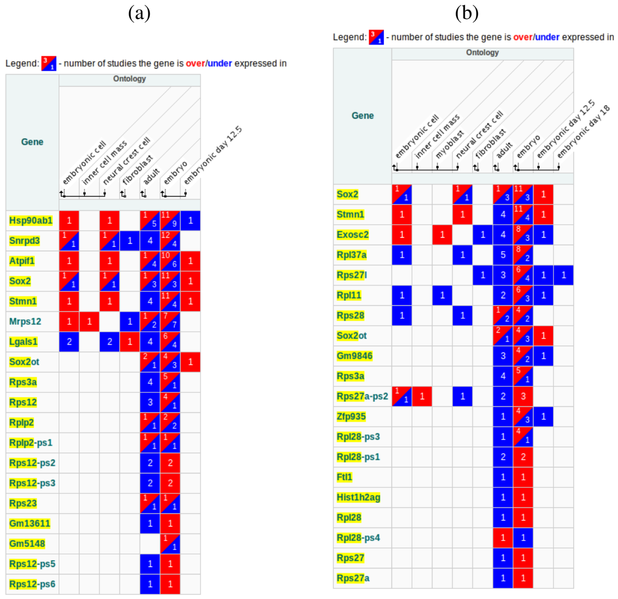


Fig. 2 (a) & (b) [(c) & (d)] PCA embeddings of samples [genes],black: mESC, red: MEF. (a) & (c) {(b) & (d)} The horizontal axis corresponds to the first {third} PC and the vertical axis corresponds to the second {fourth} PC.

For example, Sox2 is one of famous Yamanaka factors used for generation of iPS cell. Stmn1 was frequently reported to be expressive in cancer which is believed to be close to ES cell. However, most gene are ribosome related genes. Although it is reasonable, it will be more interesting for us to find that the other genes are expressive, too.

4. Conclusion

Single gene expression analysis showed that gene expression profiles in each cell is scattered, but biologically important genes in mESC are expressive. Suitable treatment of single cell RNA-seq data turns out to be informative to understand gene expression regulation during cell differentiation from ES cell.



(a)

Legend: 🟥1🟦1 - number of studies the gene is over/under expressed in

Gene	embryonic cell	inner cell mass	neural crest cell	fibroblast	adult	embryo	embryonic day 12.5
Hsp90ab1	🟥1		🟥1		🟥1 🟦5	🟥11 🟦9	🟦1
Snrpd3	🟥1 🟦1		🟥1 🟦1	🟦1	🟦4	🟥12 🟦4	
Atpif1	🟥1		🟥1		🟥1 🟦4	🟥10 🟦6	🟥1
Sox2	🟥1 🟦1		🟥1 🟦1		🟥1 🟦3	🟥11 🟦3	🟥1
Stmn1	🟥1		🟥1		🟦4	🟥11 🟦4	🟥1
Mrps12	🟥1	🟥1		🟦1	🟥1 🟦2	🟥7 🟦7	
Lgals1	🟦2		🟦2	🟥1	🟦4	🟥6 🟦4	
Sox2ot					🟥2 🟦1	🟥4 🟦3	🟥1
Rps3a					🟦4	🟥5 🟦1	
Rps12					🟦3	🟥4 🟦1	
Rplp2					🟥1 🟦1	🟥2 🟦2	
Rplp2-ps1					🟥1 🟦1	🟥1 🟦1	
Rps12-ps2					🟦2	🟥2	
Rps12-ps3					🟦2	🟥2	
Rps23					🟥1 🟦1	🟥1 🟦1	
Gm13611					🟦1	🟥1	
Gm5148						🟥1 🟦1	
Rps12-ps5					🟦1	🟥1	
Rps12-ps6					🟦1	🟥1	

(b)

Legend: 🟥1🟦1 - number of studies the gene is over/under expressed in

Gene	embryonic cell	inner cell mass	myoblast	neural crest cell	fibroblast	adult	embryo	embryonic day 12.5	embryonic day 18
Sox2	🟥1 🟦1			🟥1 🟦1		🟥1 🟦3	🟥11 🟦3	🟥1	
Stmn1	🟥1			🟥1		🟦4	🟥11 🟦4	🟥1	
Exosc2	🟥1		🟥1		🟦1	🟦4	🟥8 🟦3	🟦1	
Rpl37a	🟦1			🟦1		🟦5	🟥8 🟦2		
Rps27l					🟦1	🟦3	🟥6 🟦4	🟦1	🟦1
Rpl11	🟦1		🟦1			🟦2	🟥6 🟦3	🟦1	
Rps28	🟦1			🟦1		🟥1 🟦2	🟥4 🟦2		
Sox2ot						🟥2 🟦1	🟥4 🟦3	🟥1	
Gm9846						🟦3	🟥4 🟦2	🟦1	
Rps3a						🟦4	🟥5 🟦1		
Rps27a-ps2	🟥1 🟦1	🟥1		🟦1		🟦2	🟥3		
Zfp935						🟦1	🟥4 🟦3	🟦1	
Rpl28-ps3						🟦1	🟥4 🟦1		
Rpl28-ps1						🟦2	🟥2		
Ftl1						🟦1	🟥1		
Hist1h2ag						🟦1	🟥1		
Rpl28						🟦1	🟥1		
Rpl28-ps4						🟥1	🟦1		
Rps27						🟦1	🟥1		
Rps27a						🟦1	🟥1		

Fig. 3 Gene expression profile in gene expression atlas for genes selected by PCA.

Acknowledgments This study was supported by KAKENHI (23300357).

References

[1] Mansergh, F. C., Daly, C. S., Hurley, A. L., Wride, M. A., Hunter, S. M. and Evans, M. J.: Gene expression profiles during early differentiation of mouse embryonic stem cells, *BMC Dev. Biol.*, Vol. 9, p. 5 (2009).

[2] Seshi, B.: Gene expression analysis at the single cell level using the human bone marrow stromal cell as a model: sample preparation methods, *Methods Mol. Biol.*, Vol. 449, pp. 117–132 (2008).

[3] Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J. B., Lonnerberg, P. and Linnarsson, S.: Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq, *Genome Res.*, Vol. 21, No. 7, pp. 1160–1167 (2011).

[4] Trapnell, C., Pachter, L. and Salzberg, S. L.: TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*, Vol. 25, No. 9, pp. 1105–1111 (2009).

[5] Roberts, A., Pimentel, H., Trapnell, C. and Pachter, L.: Identification of novel transcripts in annotated genomes using RNA-Seq, *Bioinformatics*, Vol. 27, No. 17, pp. 2325–2329 (2011).

[6] Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F. and Liu, J.: MapSplice: accurate mapping of RNA-seq reads for splice junction discovery, *Nucleic Acids Res.*, Vol. 38, No. 18, p. e178 (2010).

[7] Xu, G., Deng, N., Zhao, Z., Judeh, T., Flemington, E. and Zhu, D.: SAMMate: a GUI tool for processing short read alignments in SAM/BAM format, *Source Code Biol Med*, Vol. 6, No. 1, p. 2 (2011).

[8] Kapushesky, M., Adamusiak, T., Burdett, T., Culhane, A., Farne, A., Filippov, A., Holloway, E., Klebanov, A., Kryvych, N., Kurbatova, N., Kurnosov, P., Malone, J., Melnichuk, O., Petryszak, R., Pultsin, N., Rustici, G., Tikhonov, A., Travillian, R. S., Williams, E., Zorin, A., Parkinson, H. and Brazma, A.: Gene Expression Atlas update–a value-added database of microarray and sequencing-based functional genomics experiments, *Nucleic Acids Res.*, Vol. 40, No. Database issue, pp. D1077–1081 (2012).

[9] Tusher, V. G., Tibshirani, R. and Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response, *Proc. Natl. Acad. Sci. U.S.A.*, Vol. 98, No. 9, pp. 5116–5121 (2001).

[10] Schwender, H.: *siggenes: Multiple testing using SAM and Efron's empirical Bayes approaches* (2009). R package version 1.24.0.

[11] R Development Core Team: *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2010). ISBN 3-900051-07-0.