Novel and Significant Spermatogenesis-related Gene Selection and Confirmation with Microarrays

Weixiang Liu[1] **Aifa Tang**[2] **Kehong Yuan**[3] **Datian Ye**[3]

[1]Shenzhen Key Lab of Biomedical Engineering
Shenzhen University, Shenzhen, China 518060
[2]Shenzhen Key Lab of Male Reproduction and Genetics
Peking University Shenzhen Hospital, Shenzhen, China 518036
[3]Life Science Division, Graduate School at Shenzhen
Tsinghua University, Shenzhen, 518055, China

Abstract

Microarray provides a large scale gene selection platform in post-genomics era. However biological verification of computational results is also time consuming. In this paper we apply computational methods to select novel and significant spermatogenesis-related genes based on microarray data from our research group and confirm them by other microarray dataset. Known and novel unclear genes are tested in our experiments and the consistency of expression profiles of selected genes during development from different microarray datasets demonstrates the effectiveness of the method.

Keywords: Spermatogenesis, microarray, gene expression level, gene selection, confirmation.

1. Introduction

Sperm development, termed spermatogenesis, is characterized by a mitotic (spermatogonia), a meiotic (spermatocytes) and a differentiative haploid (spermatids) phase. The understanding of the biological function of key genes during spermatogenesis is helpful for the treatment of male sterility caused by abnormity of sperm development, protecting from withering of male sexual potency, and offering new contraceptive targets and health care drugs [1]. The identification of genes encoded proteins that play specific roles in different steps of germ cell development should shed fresh light on the mechanisms analysis of spermatogenesis. The list of genes that are involved in the mammalian spermatogenesis is rapidly growing. For example, CERM [2], tesmin [3], MSJ-1 [4], Soggy [5], SP-10 [6], mTSARG3 [7], SRG4 [8] , TESF-1 [9] have been proved to be the spermatogenesis-related genes in previous study.

To know how genes are controlled to express at the exact time and which genes function uniquely in special cells during spermatogenesis, an important step is taken to define gene profiles. The advent of DNA microarray technology has offered the promise of casting new insights onto deciphering secrets of life by monitoring activities of thousands of genes simultaneously. DNA microarray is a useful, high throughput method, which provides a platform to evaluate the abundance of genes in parallel, allowing monitoring

Proceedings of the 11th Joint Conference on Information Sciences (2008)
Published by Atlantis Press

changes in gene expression during developmental events [10].

Sha et al [11] have compared gene expression profiles between adult and fetal human testes by the use of a self-made cDNA chip that comprised of 9,216 genes. 731 different expressing genes have been characterized comprising of 54 known genes, in which 18.52% were exclusively expressed in spermatogenesis. Guo et al. [12] isolated six different types of spermatogenic cells (primitive type A spermatogonia, type B spermatogonia, preleptotene spermatocytes, pachytene spermatocytes, round spermatiss and elongating spermatids) from Balb/C mice testes. Atlas cDNA arrays containing 1,176 known mouse genes were used to determine the gene expression profiles of the spermatogenic cells. The expressions of 260 genes were detected and 115 genes show differential expression in six different stages of the spermatogenic cells.

Recently, we have isolated testis from 4, 9, 18, 35, 54 days and 6 months old Balb/C mice. cRNAs prepared from these testis samples have been hybridized with commercially available GeneChip Mouse Genome 430 2.0 Array (Affymetrix Inc.) chip, which contained 34,000 known mouse genes and 8,000 unknown genes or ESTs (Expressed Sequence of Tags), and thus spanning the whole mouse genome. In mining the microarray data, we identified 2058 gradually up-regulated transcripts from four days to six months of testis samples. These transcripts, including some known and unknown genes or ESTs, should be related to mouse testis development and spermatogenesis as discussed in our previous work [13, 14, 15]. However the characteristics of some differently expressed transcripts by manual analysis and the confirmation of authenticity with semi-quantitative RT-PCR are limited to small scale data analysis and time consuming.

In this paper, basing on our microarray data, we apply computational methods to find novel and significant spermatogenesis related genes and confirm them with other microarray dataset. Our test on known and unclear genes demonstrates the capacity of large scale computational data analysis and strong correlation confirmed between microarray datasets from different research groups.

2. Materials and methods

2.1. Novel and Significant Spermatogenesis-related Gene Selection and Confirmation

In our previous work [15, 13, 14], 2058 differential expression genes are selected and some novel genes are investigated. However the characteristics of these differently expressed transcripts by manual analysis and the confirmation of authenticity with semi-quantitative RT-PCR are limited to small scale data analysis and time consuming. Here we give a further analysis on the microarray data using computational methods for large scale gene selection and confirmation by other microarray data.

2.2. Novel gene selection

GeneChip Mouse Genome 430 2.0 Array (Affymetrix Inc.) chip includes $\sim 34,000$ known mouse genes and $\sim 8,000$ unknown/unclear genes or ESTs. We selected novel genes according to the gene symbol item from the annotation file of the chip, i.e. the gene/EST is novel if there is no one gene symbol item for the probe in the annotation.

2.3. Significant Gene Selection

Here we used the max-min ratio across all samples to select significant genes where max and min correspond to the maximum and minimum expression levels of each gene. The max-min ratio is defined as

$$\mathrm{FC}_i^{mm} = \frac{\max_j g_{ij}}{\min_j g_{ij}}. \qquad (1)$$

2.4. Confirmation with Other Microarray Data

According the GeneChip Mouse Genome 430 2.0 Array, we found one microarray data [16] which contains 8 samples at 4 time courses (two samples for each time point) in the development of spermatogenesis: Type A spermatogonia (TAS), Type B spermatogonia (TBS), Pachytene spermatocytes (PS), and Round spermatids (RS). The data is available in the NCBI's Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) with the accession number GSE4193.

3. Experimental Results

For our microarray dataset as described in [15], here we set floor value as 100 and ceiling value as 16, 000 in preprocessing. The final data is a 2058 × 6 matrix. The confirmation data for same 2058 genes from the GSE4193 dataset is extracted and the average value of each gene at every time point is calculated. Then the data is formed in a 2058 × 4 matrix.

As a demonstration, we first selected 5 top genes from known genes and unknown genes, respectively based on our microarray. And then we selected same genes from the GSE4193 dataset. The selected 5 top known genes are listed in Table 1 and 5 top unknown genes in Table 2, respectively.

Table 1: The 5 top known genes. Index is the number in the 2058 gene list.

Rank	Index	Probe	Name
1	315	1421682_a_at	Tcte3
2	1024	1415924_at	Tnp1
3	1020	1448105_at	Prm2
4	25	1415846_a_at	Ldhc
5	1060	1422419_s_at	Tnp2

Table 2: The 5 top unknown genes. Index is the number in the 2058 gene list.

Rank	Index	Probe
1	4	1437098_x_at
2	158	1455921_at
3	303	1429868_at
4	20	1439402_at
5	667	1436348_at

For confirmation, all gene expression levels are shown in Figure 1. In the 5 top known genes, two genes (gene Tcte3 with index no. 315 and gene Ldhc with index no. 25) have low levels at times courses day 4 and 9 while they have high levels at and after day 18 from our microarray data as shown in Figure 1-(b). The expression levels of these two genes from the GSE4193 dataset also have similar trends with break-point at day 8 as shown in Figure 1-(d). Other 3 genes (gene Tnp1 with index no. 1024, gene Prm2 with index no. 1020 and gene Tnp2 with index no. 1060) have the break-points at day 18 for our microarray data and at day 11 for the GSE4193 dataset. In summary, the 5 top known genes have consistent expression profiles in two different microarray datasets. And this provides a confirmation for our gene selection from our microarray data. For the 5 top unknown genes, the consistency of expression profiles is also indicated from Figure 1-(a) and (c). For example, three genes (Nos: 4, 158 and 667) have high expression levels at three time points (day 35, day 54 and month 6) from our dataset and at two time points (day 11 and day 26) from the GSE4193 dataset.

In order to quantify the strength of correlation of selected genes on two different

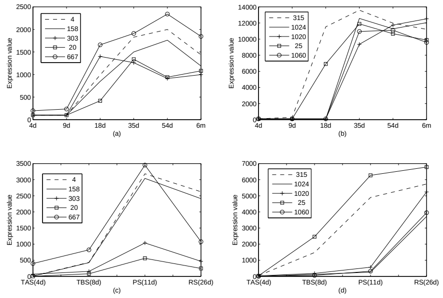

Fig. 1: Confirmation within two microarray datasets. Expression levels of 5 unknown (left) and 5 known (right) spermatogenesis related genes selected from our dataset (top) and confirmed by other one dataset (bottom) from mouse chip. Top 5 unknown genes (Nos: 4, 158, 303, 20 and 667 from our gene list) and top 5 known genes (Nos: 315, 1024, 1020, 25 and 1060 from our gene list). In (a) and (b), 6 time courses in the development of spermatogenesis are: 4, 9, 18, 35, 54 days and 6 months; in (c) and (d), 4 time courses in the development of spermatogenesis are: Type A spermatogonia (TAS; about at day 4), Type B spermatogonia (TBS; about at day 8), Pachytene spermatocytes (PS; about at day 11), and Round spermatids (RS; about at day 26).

datasets, we calculated the pairwise linear correlation coefficient matrices of known and unknown genes on two datasets, respectively. For calculation, the expression levels of the first four time points in our dataset are used. The results are shown in Tables 3 and 4, respectively. In each table, the 5×5 coefficient matrix, the (i, j)-th element reflects the linear correlation strength of i-th gene in our dataset and j-th gene in GSE4193 dataset and each diagonal element reflects the correlation strength of same gene between two datasets. From Table 3, we can see that two genes (Nos: 315 and 25) have strong correlations (with values > 0.9 underlined) between two datasets and other 3 genes (Nos: 1024, 1020 and 1060) do. And this is consistent with results of the expression profiles in Figure 1-(b) and (d). From Table 4, we can see that four genes (Nos: 4, 158, 303 and 667) have strong correlations (with values > 0.8, underlined) between two datasets. This is also consistent with the results from Figure 1 (a) and (c).

4. Conclusions

We applied computational method for gene selection and confirmation in spermatogenesis with microarray datasets. We selected significant known and unknown genes from our microarray data and confirmed them with other microarray data. The consistency of expression profiles of all selected genes within two microarray datasets demonstrates the effectiveness of the method and it provides a strategy for large scale gene selection and confirmation with microarray data.

For future work, we will consider other advanced gene selection methods and integrate more genomical data for system analysis. Furthermore, we will investigate the function of selected genes and confirm it with comprehensive experiments.

Acknowledgements

This work was supported by the Open Fund of Shenzhen Key Lab of Biomedical Engineering , and the National Natural Science Foundation of China (No. 30700824). The authors would like to thank Ms. Lu Sun and Dr. Zhiqiang Wang for help and discussion on the experimental data. The first author would also like to thank Dr. Nikolaus Schultz for his help and discussion on time course of spermatogenesis.

Table 3: The pairwise linear correlation coefficient matrix of known genes on two datasets. The value larger than 0.9 is underlined.

	315	1024	1020	25	1060
315	0.9796	0.6935	0.7249	0.9538	0.7090
1024	0.6630	0.9993	0.9968	0.6019	0.9985
1020	0.6630	0.9993	0.9968	0.6019	0.9985
25	0.9513	0.8466	0.8695	0.9111	0.8582
1060	0.6630	0.9993	0.9968	0.6019	0.9985

Table 4: The pairwise linear correlation coefficient matrix of unknown genes on two datasets. The value larger than 0.8 is underlined.

	4	158	303	20	667
4	0.8358	0.8182	0.5368	0.4692	0.3432
158	0.8454	0.8283	0.5519	0.4851	0.3600
303	0.9925	0.9893	0.8758	0.8359	0.7494
20	0.6601	0.6364	0.2911	0.2153	0.0806
667	0.9623	0.9534	0.7691	0.7168	0.6132

References

[1] N. Schultz, F.K. Hamra, and D.L. Garbers. A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. *Proceedings of the National Academy of Sciences*, 100(21):12201–12206, 2003.

[2] J.A. Blendy, K.H. Kaestner, G.F. Weinbauer, E. Nieschlag, and G. Schuetz. Severe impairment of permatogenesis in mice lacking the CREM gene. *Nature*, 380(6570):162–165, 1996.

[3] T. Sugihara, R. Wadhwa, SC Kaul, and Y. Mitsui. A Novel Testis-Specific Metallothionein-like Protein, Tesmin, Is an Early Marker of Male Germ Cell Differentiation. *Genomics*, 57(1):130–136, 1999.

[4] G. Berruti, L. Perego, B. Borgonovo, and E. Martegani. MSJ-1, a New Member of the DNAJ Family of Proteins, Is a Male Germ Cell-Specific Gene Product. *Experimental Cell Research*, 239(2):430–441, 1998.

[5] K.J. Kaneko and M.L. DePamphilis. Soggy, a spermatocyte-specific gene, lies 3.8 kb upstream of and antipodal to TEAD-2, a transcription factor expressed at the beginning of mouse development. *Nucleic Acids Research*, 28(20):3982–3990, 2000.

[6] P.P. Reddi, A.N. Shore, K.K. Acharya, and J.C. Herr. Transcriptional regulation of spermiogenesis: insights from the study of the gene encoding the acrosomal protein SP-10. *Journal of Reproductive Immunology*, 53(1-2):25–36, 2002.

[7] G. Liu, GX Lu, JJ Fu, SF Liu, and XW Xing. Molecular cloning of mTSARG3 gene related to apoptosis in mouse spermatogenic cells. *Acta Biochimica et Biophysica Sinica*, 35(12):1133–9, 2003.

[8] XW Xing, LY Li, G. Liu, JJ Fu, XJ Tan, and GX Lu. Identification of a novel gene SRG4 expressed at specific stages of mouse spermatogenesis. *Acta Biochim Biophys Sin*

(Shanghai), 36(5):351–9, 2004.

[9] J. Fan, M. Graham, H. Akabane, LL Richardson, and GZ Zhu. Identification of a novel male germ cell-specific gene TESF-1 in mice. *Biochem Biophys Res Commun*, 340:8–12, 2005.

[10] AC Pease, D. Solas, EJ Sullivan, MT Cronin, CP Holmes, and SPA Fodor. Light-Generated Oligonucleotide Arrays for Rapid DNA Sequence Analysis. *Proceedings of the National Academy of Sciences*, 91(11):5022–5026, 1994.

[11] J. Sha. Identification of testis development and spermatogenesis-related genes in human and mouse testes using cDNA arrays. *Molecular Human Reproduction*, 8(6):511–517, 2002.

[12] R. Guo, Z. Yu, J. Guan, Y. Ge, J. Ma, S. Li, S. Wang, S. Xue, and D. Han. Stage-specific and tissue-specific expression characteristics of differentially expressed genes during mouse spermatogenesis. *Molecular Reproduction and Development*, 67(3):264–272, 2004.

[13] A. Tang, Z. Yu, Y. Gui, H. Zhu, Y. Long, and Z. Cai. Identification of a novel testis-specific gene in mice and its potential roles in spermatogenesis. *Croat Med J*, 48(1):43–50, 2007.

[14] Z. Yu, A. Tang, Y. Gui, X. Guo, H. Zhu, Y. Long, Z. Li, and Z. Cai. Identification and characteristics of a novel testis-specific gene, Tsc21, in mice and human. *Molecular Biology Reports*, 34(2):127–134, 2007.

[15] A. Tang, Z. Yu, Y. Gui, X. Guo, Y. Long, and Z. Cai. Identification and Characteristics of a Novel Testis-Specific Gene, Tsc24, in Human and Mice. *Biological & Pharmaceutical Bulletin*, 29(11):2187–2191, 2006.

[16] S.H. Namekawa, P.J. Park, L.F. Zhang, J.E. Shima, J.R. McCarrey, M.D. Griswold, and J.T. Lee. Post-meiotic Sex Chromatin in the Male Germline of Mice. *Current Biology*, 16(7):660–667, 2006.