Data in Brief

LncRNA analysis of mouse spermatogonial stem cells following glial cell-derived neurotrophic factor treatment

Lufan Li, Min Wang, Mei Wang, Xiaoxi Wu, Lei Geng, Yuanyuan Xue, Xiang Wei, Yuanyuan Jia, Xin Wu *

State Key Laboratory of Reproductive Medicine (SKLRM), Nanjing Medical University, Nanjing, Jiangsu 210029, China

A B S T R A C T

Spermatogonial stem cells (SSCs) are the foundation of spermatogenesis. Long non-coding RNAs (lncRNAs) are a class of non-coding RNAs with at least 200 bp in length, which play important roles in various biological processes. Growth factor glial cell line-derived neurotrophic factor (GDNF), secreted from testis niches, is critical for self-renewal of SSCs in vitro and in vivo. Using Illumina HiSeq™ 2000 high throughput sequencing, we found 55,924 lncRNAs which were regulated by GDNF in SSCs in vitro; these included 21,929 known lncRNAs from NONCODE library (version 3.0) and 33,975 predicted lncRNAs which were identified using Coding Potential Calculator. Analyses of these data should provide new insights into regulated mechanism in SSC self-renewal and proliferation. The data have been deposited in the Gene Expression Omnibus (series GSE66998).

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66998.

2. Experimental design, materials and methods

2.1. Experimental design

SSCs were cultured in a medium supplemented with growth factor GDNF, the essential cytokine supporting cell maintenance and expansion in vitro [1–3]. SSC samples were collected from two independent cell lines in three different culture conditions, including normal culture in GDNF and FGF2 supplemented medium, after 18 h of GDNF depletion, 8 h of GDNF replenishment, and 18 h of GDNF withdrawal. After RNA was isolated and processed, the lncRNA expression profiling was detected and analyzed using Illumina HiSeq™ 2000, and followed by the analysis and annotation of sequencing data using commercial services (BGI). Detailed experimental procedure for cell treatment was shown in Fig. 1, RNA processing for sequencing was shown in Fig. 2, and data bioinformatics analysis was shown in Fig. 3.

2.2. Materials and methods

2.2.1. SSC culture and RNA isolation

SSCs were isolated from 8 d old mouse testis using magnetic-activated cell sorting (MACS) isolation for THY1-positive (CD90.2) cells, as previously described [4,5]. Long-term SSC self-renewal and proliferation were supported in a chemically defined, serum-free minimal essential medium alpha (MEM a) medium (mSFM) supplemented with 20 ng/ml of GDNF (R&D Systems), 150 ng/ml of GFRA1 (R&D Systems), and 1 ng/ml of basic fibroblast growth factor (FGF2; BD Biosciences) at 37 °C. The medium was replaced every 2–3 days and...
cells were sub-cultured at approximately 7-day intervals. RNA was isolated from individual culture according to standard Trizol isolation protocols. RNA with an A260:A280 ratio of 1.8 or greater was applied for further sequencing.

2.2.2. RNA processing, sequencing and bioinformatics analysis

mRNA and non-coding RNAs extracted from total RNA were first enriched by removing rRNA. The mRNAs and non-coding RNAs were then fragmented into about 200–500 nt in fragmentation buffers. The first-strand cDNA was synthesized by a random hexamer-primer using the fragments as templates, and dTTP was substituted by dUTP during the synthesis of the second strand. Short fragments were purified and resolved with EB buffer for end reparation and single nucleotide A (adenine) addition. After that, the short fragments were connected with adapters, then the second strand was degraded finally using UNG (Uracil-N-Glycosylase) [6]. After agarose gel electrophoresis, the suitable fragments were selected for the PCR amplification as templates. During the QC steps, Agilent 2100 Bioanaylzer and ABI StepOnePlus Real-Time PCR System were applied in quantification and qualification of the sample library. At last, the library was subjected to Illumina HiSeq™ 2000 sequencing.

The original image data was transferred into sequence data via base calling, which was defined as raw data or raw reads. Before doing any further analysis, quality control was required in order to detect whether the data was qualified. In addition, filtering of raw data was needed to decrease data noise. Filtering steps were as follows: 1) Remove reads with adapters; 2) remove reads in which unknown bases are more than 10%; and 3) remove low quality reads (the percentage of low quality bases is over 50% in a read, we defined the low quality base to be the base whose sequencing quality is no more than 10). After filtering, the remaining reads were called “clean reads” and used for downstream bioinformatics analysis. Effected by the sample quality and species, the rRNA removing efficiency could be instable, while rRNA pollution would interfere further analysis, therefore clean reads were mapped to rRNA reference using short read alignment software SOAPaligner/ SOAP2 [7] to remove the remain rRNA reads, and the reads left were used to proceed transcriptome assembling and quantification.

Transcript assembling: First, rRNA removed reads were mapped to reference genome using an improved version of TopHat2 [8], which can align reads across splice junction without relying on gene
annotation. Then, reads mapped to genome will be assembled by Cufflinks [9]. We performed Reference Annotation Based Transcript (RABT) [10] assembly with the reference gene annotation to compensate incompletely assembled transcripts caused by read coverage gaps in the regions of reference gene. The set of transfrags generated was then compared with the reference transcripts to remove transfrags that were approximately equivalent to the whole or a portion of a reference transcript.

After the assembling, we obtained the whole parsimonious set of transcripts. Then, we compared the assembled transcripts to the reference annotation by utilizing Cuffcompare [9]. We blasted these transcripts with the NONCODE ncRNA library (http://noncode.org) using the filter set (identity > 0.9 and coverage > 0.8) to identify the known long non-coding RNA. For those lncRNA were not mapping to NONCODE ncRNA library, we evaluated and compared several software for lncRNA prediction, and choose Coding Potential Calculator (CPC) [11], which perform well across other software in both accuracy and efficiency. A true protein-coding transcript is more likely to have a long and high-quality Open Reading Frame (ORF) compared with a non-coding transcript. First three features assess the extent and quality of the ORF in a transcript. LOG-ODDS SCORE: an indicator of the quality of a predicted ORF. COVERAGE OF THE PREDICTED ORF: an indicator of ORF quality. INTEGRITY OF THE PREDICTED ORF: whether an ORF begins with a start codon and ends with an in-frame stop codon. Another three features from parsing the output of BLASTX: NUMBER OF HITS: a true protein-coding transcript is likely to have more hits with known proteins than a non-coding transcript does, HIT SCORE: a true protein-coding transcript where the hits are also likely to have higher quality; i.e. the HSPs (High-scoring Segment Pairs) overall tend to have lower E-value, and FRAME SCORE: to measure the distribution of the HSPs among three reading frames. Based on the six features, we could obtain the novel long non-coding RNA. LncRNA expression analyses were performed by BIG Company (Shenzhen, BIG Company, China) as reported previously [12]; P values correspond to differential gene expression tests and FDR (False Discovery Rate) was used to determine the threshold of P-value in multiple tests.

### 3. Data summary

The preliminary data from analysis for LncRNA were summarized in Supplemental Tables 1 & 2 (excel format). We found 55,924 LncRNAs which were presented in SSCs in vitro; these included 21,929 known LncRNAs from NONCODE library (version 3.0) and 33,975 predicted LncRNAs which were identified using Coding Potential Calculator. A comparison of RNA species with at least a two-fold change (P < 0.05) in the expression level between Normal (N), 18 h GDNF withdrawal (0 h) and 8 h GDNF re-exposure (8 h) identified 805 LncRNAs. More stringent criteria (false detection rate, FDR < 0.05) identified 83 LncRNAs that exhibited distinct expression patterns following GDNF treatment, which is an essential growth factor required for SSC self-renewal. These represent approximately 1.4% of the total sequenced known and predicted lncRNAs, suggesting a potential biological significance of these LncRNAs in GDNF dependent self-renewal of SSCs: Supplementary Table 1: normal cultured SSCs with 18 h GDNF withdrawal SSCs, the differential expression of LncRNAs, Supplementary Table 2: 8 h GDNF refreshed cultured SSCs with 18 h GDNF withdrawal SSCs, the differential expression of LncRNAs.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2015.06.012.

### Competing financial interests

The authors declare no competing financial interests.

### Funding supports

This study was supported by the National Natural Science Foundation of the People’s Republic of China Grant 31271581, National Key Basic Research Program Grant (973 Grant, 2013CB945201), Natural Science Foundation of Jiangsu Province (BK2012837) and Research Foundation for the Doctoral Program of Higher Education of China (20122334120013).

### References

[1] M. Kanatsu-Shinohara, N. Ogouni, K. Inoue, H. Miki, A. Ogura, S. Toyokuni, T. Shinohara, Long-term proliferation in culture and germline transmission of mouse male germ-line stem cells. Biol. Reprod. 69 (2003) 612–616.
[2] H. Kubota, M.R. Avarbock, R.L. Brinster, Growth factors essential for self-renewal and expansion of mouse spermatogonial stem cells. Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 16489–16494.
[3] X. Meng, M. Lindahl, M.E. Hyvönen, M. Parvinen, D.G. de Rooij, M.W. Hess, A.Raatto, H. Sariola, Regulation of cell fate decision of undifferentiated spermatogonia by GDNF. Science 287 (2000) 1489–1493.
[4] J.M. Oatley, M.R. Avarbock, A.I. Telaranta, D.T. Fearon, R.L. Brinster, Identifying genes important for spermatogonial stem cell self-renewal and survival. Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 9524–9529.
[5] H. Kubota, M.R. Avarbock, R.L. Brinster, Culture conditions and single growth factors affect fate determination of mouse spermatogonial stem cells. Biol. Reprod. 71 (2004) 722–731.
[6] D. Parkhomchuk, T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach, A. Soldatov, Transcriptome analysis by strand-specific sequencing of complementary DNA, Nucleic Acids Res. 37 (2009) e123, 1-7.
[7] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yu, K. Kristiansen, J. Wang, SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 15 (2009) 1966–1967.
[8] D. Kim, G. Perteza, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 4 (2013) R36.
[9] C. Trapnell, B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold, L. Pachter, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 5 (2010) 511–515.

[10] A. Roberts, H. Pimentel, C. Trapnell, L. Pachter, Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics 17 (2011) 2325–2329.

[11] L. Kong, Y. Zhang, Z.Q. Ye, X.Q. Liu, S.Q. Zhao, L. Wei, G. Gao, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. 35 (Suppl. 2) (2007) W345–W349.

[12] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, A. Conesa, Differential expression in RNA-seq: a matter of depth. Genome Res. 21 (2011) 2213–2223.