Genomics Data 5 (2015) 366–367

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Molecular subtyping of leiomyosarcoma with 3′ end RNA sequencing

Xiangqian Guo [a,b], Erna Forgó [a], Matt van de Rijn [a,*]

[a] Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305, USA
[b] Department of Biochemistry and Molecular Biology, Medical School of Henan University, Kaifeng, Henan 475004, China

ARTICLE INFO

Article history:
Received 19 June 2015
Accepted 24 June 2015
Available online 9 July 2015

Keywords:
Leiomyosarcoma
Subtypes
Expression profiling
3′ end RNA sequencing

ABSTRACT

Leiomyosarcoma (LMS) is a malignant neoplasm with smooth muscle differentiation. Little is known about its molecular heterogeneity and no targeted therapy currently exists for LMS. We performed expression profiling on 99 cases of LMS with 3′ end RNA sequencing (3SEQ) and demonstrated the existence of 3 molecular subtypes in this cohort. We consequently showed that these molecular subtypes are reproducible using an independent cohort of 82 LMS cases from TCGA. Two new formalin-fixed, paraffin-embedded (FFPE) tissue-compatible diagnostic immunohistochemical markers were identified for two of the three subtypes: LMOD1 for subtype I LMS and ARL4C for subtype II LMS. Subtype I LMS and subtype II LMS were associated with good and poor prognosis, respectively. Here, we describe the details of LMS diagnosis, RNA isolation, 3SEQ library construction, 3SEQ sequencing data analysis and molecular subtype determination. The 3SEQ data produced in this study was deposited into Gene Expression Omnibus (GEO) under GSE45510.

Specifications	
Organism/cell line/tissue	Leiomyosarcoma, FFPE tissues, human
Sex	Male or female
Sequencer or array type	3′ end RNA sequencing
Data format	TPM (transcripts per million reads) normalized matrix
Experimental factors	Archival FFPE blocks for 99 cases of leiomyosarcoma
Experimental features	Total RNA isolation, oligo(dT) selection and gene expression profiling of 99 leiomyosarcomas
Consent	IRB approval and a waiver of consent due to the archival nature of the specimens
Sample source location	Nine hospitals from United States, Canada and Europe (see Experimental design)

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45510.

* Corresponding author at: Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA.
E-mail address: mrijn@stanford.edu (M. van de Rijn).

2. Experimental design, materials and methods

2.1. Experimental design

To explore the molecular subtypes of leiomyosarcoma (LMS), paraffin blocks of 99 LMS cases from 1991 to 2012 from nine hospitals (Stanford Hospital, Brigham and Women's Hospital, McKay-Dee Hospital Center, St. Luke's Hospital, Baptist Health Medical Center, Ingalls Hospital, Vancouver General Hospital, Hospital de la Santa Creu i Sant Pau and Alta Bates Summit Medical Center), were collected with IRB approval and a waiver of consent due to the archival nature of the specimens. The total RNA was extracted for these cases and subsequently analyzed by 3′ end RNA sequencing. Consensus Clustering was used to determine the optimal number of subtypes and Silhouette analysis was then performed to measure the confidence of subtype assignment per case. To test the reproducibility of molecular subtype classification from the 3SEQ dataset, the expression profiles (RNASeq data) of 82 additional LMS cases were downloaded from The Cancer Genome Atlas (TCGA) database and analyzed in an identical way to the 3SEQ data. Subclass mapping was then used to find the common subtypes identified in both datasets [1].

2.2. Materials

The LMS cases used in this study were formalin-fixed, paraffin-embedded (FFPE) tissues.

http://dx.doi.org/10.1016/j.gdata.2015.06.029

2.3. 3SEQ library construction

After LMS FFPE blocks were obtained, two experienced pathologists (one from Stanford University and one from Brigham and Women's Hospital) assessed and circled the regions comprised of LMS tumor cells. Samples with paucity of material or poor preservation of material were excluded. Multiple 2 mm-diameter cores from the areas circled were re-embedded longitudinally into new paraffin blocks, and were sectioned and re-evaluated for the second time by H&E staining to ensure the purity of the samples. Only cores with ≥90% of tumor cells were processed for subsequent RNA extraction.

The total RNA of the LMS cases was extracted using RecoverAll™ Total Nucleic Acid Isolation kit (Ambion, Cat # 1975). The quality of total RNA was assessed by agarose gel electrophoresis and used to determine the amount of time necessary for shearing of the total RNA by heat in first strand buffer (Invitrogen, Cat # 18080-044) for subsequent 3SEQ library construction. The 3SEQ library construction included the following steps; first strand cDNA synthesis with Superscript III Reverse Transcriptase (Invitrogen, Cat # 18080-044), second strand cDNA synthesis with *E coli* DNA ligase (Invitrogen, Cat # 18052-019) and *E. coli* DNA polymerase I (New England Biolab, Cat # M0209L), followed by the addition of 'A' to the 3′ end of double strand DNA fragments with Klenow exo (3′ to 5′ exo minus, New England Biolab, Cat # M0212L), ligation of adapters (Illumina, Cat # 1001782 —OLIGO MIX) with DNA ligase (New England Biolab, Cat # M2200L) and PCR amplification using 2× Phusion PCR master mix (New England Biolab, Cat # F-531S) [1–6]. The detailed protocol of 3SEQ library construction can be accessed using the following link, http://med.stanford.edu/labs/vanderijn-west/documents/3endRNAseqlibraryconstruction_update_11_4_2014.doc.

The 3SEQ libraries were sent to the Stanford Center for Genomics and Personalized Medicine to be sequenced directionally (36 bp) from 5′ end of mRNA fragments towards their poly(A) ends using Illumina GA IIx and HiSeq 2000 machines (Illumina, Inc., San Diego, CA, USA). The gene expression profiling data (3SEQ) have been deposited in the Gene Expression Omnibus (GEO) and are publicly accessible through GSE45510.

2.4. 3SEQ data analysis

Sequence reads (fastq format), first filtered for read quality, were re-filtered by fastx (fastx_artifacts_filter-v-Q 33, http://hannonlab.cshl.edu/fastx_toolkit/index.html), and mapped to the transcriptome (refMrna, downloaded from the UCSC genome browser, http://www.genome.ucsc.edu/) using SOAP2, allowing at most two mismatches [7]. The total numbers of sequence reads for each gene symbol from the transcriptome mapping were determined and used to create the gene-expression profile matrix (22,144 genes). Read counts from each library were normalized to transcripts per million reads (TPM). A custom Perl script was used to run the 3SEQ data processing and is publicly available [3].

2.5. Subtype determinant and validation

To determine the optimal number of molecular subtypes of leiomyosarcoma [1], the expression matrix of genes with the most variant expression levels, filtered with a standard deviation greater than 100 across all 99 LMS cases (1300 genes), were transformed by log2 and gene-based centering [8]. Consensus Clustering (R package ConsensusClusteringPlus) [9] was performed. This analysis was run over 1000 iterations with the settings of "Distance — (1 — Pearson correlation), 80% sample resampling, 80% gene resampling, maximum evaluated k of 12, and agglomerative hierarchical clustering algorithm". Based on the analysis we chose the optimal number of subtypes as three. Expression profiling data of the 82 additional LMS cases by RNASeq was downloaded from the TCGA database. To compare with 3SEQ data, the TCGA data were normalized into TPM and analyzed with ConsensusClusteringPlus, as was done for the 3SEQ data.

To measure the reproducibility of the LMS subtypes, the cases from both datasets (3SEQ and TCGA RNASeq) were evaluated using Silhouette analysis [10], where an LMS case was defined as a "core case" upon assignment of a positive Silhouette value. Subclass mapping was performed to determine the common LMS subtypes based on these "core cases" identified in the 3SEQ and TCGA RNASeq datasets.

In order to discover subtype-specific genes, SAMSeq [11] was performed on both datasets (3SEQ and TCGA RNASeq) between each subtype and all other subtypes with a FDR of 0.05, and significantly differentially expressed genes from the SAMSeq analysis was referred to identify the diagnostic biomarker for each LMS subtype.

Acknowledgments

This study was supported by the National Institutes of Health (Grant No. CA112270).

References

[1] X. Guo, V.Y. Jo, A. Mills, S. Zhu, C.H. Lee, I. Espinosa, M.R. Nucci, S. Varma, E. Forgo, T. Hastie, S. Anderson, K. Ganjoo, A.H. Beck, R. West, C. Fletcher, M. van de Rijn, Clinically relevant molecular subtypes in leiomyosarcoma. Clin. Cancer Res. (2015).
[2] C.H. Lee, R.H. Ali, M. Rouzbahman, A. Marino-Enriquez, M. Zhu, X. Guo, A.L. Brunner, S. Chiang, S. Leung, N. Nelnyk, D.G. Huntsman, C. Blake Gilks, T.O. Nielsen, P. Dal Cin, M. van de Rijn, E. Oliva, J.A. Fletcher, M.R. Nucci, Cyclin D1 as a diagnostic immunomarker for endometrial stromal sarcoma with YWHAE-FAM22 rearrangement. Am. J. Surg. Pathol. 36 (10) (2012) 1562–1570.
[3] X. Guo, S.X. Zhu, A.L. Brunner, M. van de Rijn, R.B. West, Next generation sequencing-based expression profiling identifies signatures from benign stromal proliferations that define stromal components of breast cancer. Breast Cancer Res. 15 (6) (2013) R117.
[4] A.L. Brunner, A.H. Beck, B. Edris, R.T. Sweeney, S.X. Zhu, R. Li, K. Montgomery, S. Varma, T. Gilks, X. Guo, J.W. Foley, D.M. Witten, C.P. Giacomini, R.A. Flynn, J.R. Pollack, R. Tibshirani, H.Y. Chang, M. van de Rijn, R.B. West, Transcriptional profiling of lncRNAs and novel transcribed regions across a diverse panel of archived human cancers. Genome Biol. 13 (8) (2012) R75.
[5] A.L. Brunner, J. Li, X. Guo, R.T. Sweeney, S. Varma, S.X. Zhu, R. Li, R. Tibshirani, R.B. West, A shared transcriptional program in early breast neoplasias despite genetic and clinical distinctions. Genome Biol. 15 (5) (2014) R71.
[6] C.H. Lee, W.B. Ou, A. Marino-Enriquez, M. Zhu, M. Mayeda, Y. Wang, X. Guo, A.L. Brunner, F. Amant, C.A. French, R.B. West, J.N. McAlpine, C.B. Gilks, M.B. Yaffe, L.M. Prentice, A. McPherson, S.J. Jones, M.A. Marra, S.P. Shah, M. van de Rijn, D.G. Huntsman, P. Dal Cin, M. Debiec-Rychter, M.R. Nucci, J.A. Fletcher, 14-3-3 fusion oncogenes in high-grade endometrial stromal sarcoma. Proc. Natl. Acad. Sci. U. S. A. 109 (3) (2012) 929–934.
[7] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen, J. Wang, SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25 (15) (2009) 1966–1967.
[8] M.J. de Hoon, S. Imoto, J. Nolan, S. Miyano, Open source clustering software. Bioinformatics 20 (9) (2004) 1453–1454.
[9] M.D. Wilkerson, D.N. Hayes, ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 26 (12) (2010) 1572–1573.
[10] R.J. Peter, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20 (1987) 13.
[11] J. Li, R. Tibshirani, Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. Stat. Methods Med. Res. (2011).