Data Article

RNA-seq data of invasive ductal carcinoma and adjacent normal tissues from a Korean patient with breast cancer

Ji Hyung Hong\textsuperscript{a}, Yoon Ho Ko\textsuperscript{a,b}, Keunsoo Kang\textsuperscript{c}

\textsuperscript{a} Division of Oncology, Department of Internal Medicine, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea
\textsuperscript{b} Cancer Research Institute, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea
\textsuperscript{c} Department of Microbiology, College of Natural Sciences, Dankook University, Cheonan 31116, Republic of Korea

**Article history:**
Received 8 February 2018
Received in revised form 28 February 2018
Accepted 16 March 2018
Available online 21 March 2018

**Keywords:**
Breast cancer
Luminal B subtype
Invasive ductal carcinoma
RNA-seq
Korean

**Abstract**

Invasive ductal carcinoma is the most common type of breast cancer. Here, we provide a whole transcriptome shotgun sequencing (called RNA-seq) dataset conducted with ten samples of invasive ductal carcinoma tissue and three samples of adjacent normal tissue from a single Korean breast cancer patient (luminal B subtype). Differentially expressed genes (DEGs) were identified with a false discovery rate (FDR)-adjusted p-value of 0.05. Gene ontology analysis identified several key pathways, including lymphocyte activation. A list of differentially expressed genes is provided. The raw data was uploaded to the sequence read archive (SRA) database and the BioProject ID is PRJNA432903.

© 2018 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

**Specifications Table**

| Subject area | Biology |
|--------------|---------|
| More specific subject area | NGS, Transcriptomics, Cancer biology |
| Type of data | Transcriptome data |

\textsuperscript{E-mail address:} kongk1204@dankook.ac.kr (K. Kang)

https://doi.org/10.1016/j.dib.2018.03.079

2352-3409 © 2018 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
**How data was acquired**

High-throughput sequencing using Illumina HiSeq2500

**Data format**

Raw (fastq)

**Experimental factors**

Breast cancer (invasive ductal carcinoma; luminal B subtype) and adjacent normal tissues

**Experimental features**

Poly(A) RNA was purified from 1 g total RNA from each sample, and cDNA was synthesized using SuperScript II (Invitrogen). Sequencing libraries were prepared using the TruSeq RNA Library preparation kit (Illumina)

**Data source location**

Seoul, Republic of Korea

**Data accessibility**

Raw data can be accessed at NCBI SRA (BioProject ID: PRJNA432903) (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA432903).

---

**Value of the data**

- This RNA-seq data provides a deep sequencing of ten samples of invasive ductal carcinoma tissue and three samples of adjacent normal tissue from a Korean breast cancer patient (luminal B subtype)
- The heterogeneous expression data from spatially distinct tumor samples can be used for various evaluation purposes.
- Gene ontology analysis revealed that lymphocyte activation and PPAR signaling pathway are significantly up- and down-regulated pathways, respectively, in breast cancer tissue compared to adjacent normal tissue.

---

1. **Data**

Total RNA was extracted from ten samples of cancer tissue (invasive ductal carcinoma; luminal B subtype) and three samples of adjacent normal tissue from a Korean patient with breast cancer. RNA-seq was performed to profile transcriptomes of breast cancer and normal samples. Differentially expressed genes were identified with an FDR-adjusted p-value cutoff of 0.05. Gene ontology analysis indicated that several pathways are associated with the onset or progression of breast cancer.

---

2. **Experimental design, materials and methods**

2.1. **RNA-seq**

One tissue sample of invasive ductal carcinoma (luminal B subtype) from breast tissue and a corresponding adjacent normal tissue were biopsied from a Korean woman with informed consent.

| Cancer | TIN score (median) | Normal | TIN score (median) |
|--------|--------------------|--------|--------------------|
| C0     | 79.6               | N1     | 79.3               |
| C1     | 80.1               | N2     | 65.6               |
| C2     | 79.7               | N3     | 67.8               |
| C3     | **59.6**           |        |                    |
| C4     | 79.1               |        |                    |
| C5     | 79.8               |        |                    |
| C6     | 80.6               |        |                    |
| C7     | 78.1               |        |                    |
| C8     | 80.1               |        |                    |
| C9     | 80.1               |        |                    |
This study was approved by the institutional review board of Catholic Medical Center (approval no. UC17TISI0015). The tumor and adjacent normal tissues were divided into ten and three samples, respectively. Poly(A) RNA was purified from 1 g total RNA from each sample, and cDNA was synthesized using SuperScript II (Invitrogen). Sequencing libraries were prepared using the TruSeq RNA Library preparation kit (Illumina) and sequenced using HiSeq. 2500 (Illumina).

2.2. RNA-seq analysis

Sequenced reads were trimmed using Trim Galore (version 0.4.2; https://www.bioinformatics. babraham.ac.uk/projects/trim_galore/) with Cutadapt (version 1.1.2) [1]. Trimmed reads were mapped to the reference human genome (hg38) using STAR (version 2.5.2b) [2]. The PCR-duplicate removal of mapped reads was performed using Sambamba (version 0.6.5) [3]. The quality of RNA-seq data was determined using RSeQC (version 2.6.4) with the transcript integrity number (TIN) score (Table 1) [4]. The abundances of RefSeq genes were estimated using Cufflinks with the Cuffnorm function (version 2.2.1) (Supplementary Table 1) [5].
2.3. Identification of differentially expressed genes

Differentially expressed genes (DEGs) between cancer and normal samples were identified using Cufflinks with the Cuffdiff function (version 2.2.1) [5]. DEGs were defined as the genes with FDR-adjusted p-values < 0.05. A total of 2456 up-regulated and 2601 down-regulated genes were identified in cancer samples compared to adjacent normal samples (Supplementary Table 2). When the low-quality RNA-seq data (C3) was excluded for DEG analysis, a total of 3199 up-regulated and 3422 down-regulated genes were identified as DEGs (Fig. 1 and Supplementary Table 3).

2.4. Gene ontology analysis

Gene ontology (GO) analysis was performed to identify key pathways regarding the DEGs that were identified without the C3 sample. The top 100 up-regulated (or down-regulated) DEGs that were highly expressed (> 10 average FPKM) were analyzed using Metascape (http://metascape.org) [6]. The GO analysis revealed that the majority of up-regulated genes were significantly associated with lymphocyte activation and that some down-regulated genes were involved in PPAR signaling pathway (Fig. 2).

Acknowledgements

This study was supported by a grant from the National R&D Program for Cancer Control, Ministry of Health & Welfare, Republic of Korea (1720100).

Transparency document. Supporting information

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.03.079.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2018.03.079.

References

[1] M. Marcel, Cutadapt removes adapter sequences from high-throughput sequencing reads, Bioinforma. Action 17 (2012) 10–12.
[2] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (2013) 15–21.
[3] A. Tarasov, A.J. Vilella, E. Cuppen, I.J. Nijman, P. Prins, Sambamba: fast processing of NGS alignment formats, Bioinformatics 31 (2015) 2032–2034.
[4] L. Wang, S. Wang, W. Li, RSeQC: quality control of RNA-seq experiments, Bioinformatics 28 (2012) 2184–2185.
[5] C. Trapnell, D.G. Hendrickson, M. Sauvageau, L. Goff, J.L. Rinn, L. Pachter, Differential analysis of gene regulation at transcript resolution with RNA-seq, Nat. Biotechnol. 31 (2013) 46–53.
[6] S. Tripathi, M.O. Pohl, Y. Zhou, A. Rodriguez-Frandsen, G. Wang, D.A. Stein, H.M. Moulton, P. De Jesus, J. Che, L.C. Mulder, E. Yanez, D. Andenmatten, L. Pache, B. Manicassamy, R.A. Albrecht, M.G. Gonzalez, Q. Nguyen, A. Brass, S. Elledge, M. White, S. Shapira, N. Hacothen, A. Karlas, T.F. Meyer, M. Shales, A. Gatorano, J.R. Johnson, G. Jang, T. Johnson, E. Verschuere, D. Sanders, N. Krogan, M. Shaw, R. Konig, S. Stertz, A. Garcia-Sastre, S.K. Chanda, Meta- and orthogonal integration of influenza "OMICs" data defines a role for UBR4 in virus budding, Cell Host Microbe 18 (2015) 723–735.