PeerJ, 2017, vol.2017, N3

Effect of method of deduplication on estimation of differential gene expression using RNA-seq

Klepikova A., Kasianov A., Chesnokov M., Lazarevich N., Penin A., Logacheva M.
Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

© 2017 Klepikova et al. Background. RNA-seq is a useful tool for analysis of gene expression. However, its robustness is greatly affected by a number of artifacts. One of them is the presence of duplicated reads. Results. To infer the influence of different methods of removal of duplicated reads on estimation of gene expression in cancer genomics, we analyzed paired samples of hepatocellular carcinoma (HCC) and non-tumor liver tissue. Four protocols of data analysis were applied to each sample: processing without deduplication, deduplication using a method implemented in samtools, and deduplication based on one or two molecular indices (MI).Wealso analyzed the influence of sequencing layout (single read or paired end) and read length. We found that deduplication without MI greatly affects estimated expression values; this effect is the most pronounced for highly expressed genes. Conclusion. The use of unique molecular identifiers greatly improves accuracy of RNA-seq analysis, especially for highly expressed genes. We developed a set of scripts that enable handling of MI and their incorporation into RNA-seq analysis pipelines. Deduplication without MI affects results of differential gene expression analysis, producing a high proportion of false negative results. The absence of duplicate read removal is biased towards false positives. In those cases where using MI is not possible, we recommend using paired-end sequencing layout.

http://dx.doi.org/10.7717/peerj.3091

Keywords

Cancer genomics, Deduplication, Differential expression, Hepatocarcinoma, RNA-seq

References

[1] Abdelgawad IA, Radwan NH, Hassanein HR. 2016. KIAA0101 mRNA expression in the peripheral blood of hepatocellular carcinoma patients: association with some clinicopathological features. Clinical Biochemistry 49:787-791 DOI 10.1016/j.clinbiochem.2015.12.016.

[2] Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology 12:R18 DOI 10.1186/gb-2011-12-2-r18.

[3] Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biology 11:R106 DOI 10.1186/gb-2010-11-10-r106.

[4] Anon. 2016. Illumina technical note`Optimizing cluster density on illumina sequencing systems''. Available at http:// support.illumina.com/ content/dam/illumina-marketing/ documents/ products/ other/ miseq-overclustering-primer-770-2014-038.pdf (accessed on 6 January 2017).

[5] Balzer S, Malde K, Grohme MA, Jonassen I. 2013. Filtering duplicate reads from 454 pyrosequencing data. Bioinformatics 29:830-836 DOI 10.1093/bioinformatics/btt047.

[6] Boshart M, Weih F, Nichols M, Schütz G. 1991. The tissue-specific extinguisher locus TSE1 encodes a regulatory subunit of cAMP-dependent protein kinase. Cell 66:849-859 DOI 10.1016/0092-8674(91)90432-X.

[7] Burriesci MS, Lehnert EM, Pringle JR. 2012. Fulcrum: condensing redundant reads from high-throughput sequencing studies. Bioinformatics 28:1324-1327 DOI 10.1093/bioinformatics/bts123.

[8] Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. 2016. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nature Reviews Genetics 17:257-271.

[9] Christopher KJ, Wang B, Kong Y, Weatherbee SD. 2012. Forward genetics uncovers Transmembrane protein 107 as a novel factor required for ciliogenesis and Sonic hedgehog signaling. Developmental Biology 368:382-392 DOI 10.1016/j.ydbio.2012.06.008.

[10] Dabney J, Meyer M. 2012. Length and GC-biases during sequencing library amplifica- tion: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. BioTechniques 52(2):87-94.

[11] Dozmorov MG, Adrianto I, Giles CB, Glass E, Glenn SB, Montgomery C, Sivils KL, Olson LE, Iwayama T, Freeman WM, Lessard CJ, Wren JD. 2015. Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data. BMC Bioinformatics 16(Suppl 13):S10 DOI 10.1186/1471-2105-16-S13-S10.

[12] Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 30:2503-2505 DOI 10.1093/bioinformatics/btu314.

[13] Flores IL, Kawahara R, Miguel MCC, Granato DC, Domingues RR, Macedo CC, Carnielli CM, Yokoo S, Rodrigues PC, Monteiro BV, Oliveira CE, Salmon CR, Nociti FH, Lopes MA, Santos-Silva A, Winck FV, Coletta RD, Paes Leme AF. 2016. EEF1D modulates proliferation and epithelial-mesenchymal transition in oral squamous cell carcinoma. Clinical Science 130:785-799 DOI 10.1042/CS20150646.

[14] Fu GK, Hu J, Wang P-H, Fodor SPA. 2011. Counting individual DNA molecules by the stochastic attachment of diverse labels. Proceedings of the National Academy of Sci- ences of the United States of America 108:9026-9031 DOI 10.1073/pnas.1017621108.

[15] Fu GK, Xu W, Wilhelmy J, Mindrinos MN, Davis RW, Xiao W, Fodor SPA. 2014. Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. Proceedings of the National Academy of Sciences of the United States of America 111:1891-1896 DOI 10.1073/pnas.1323732111.

[16] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signaling 6(269):pl1 DOI 10.1126/scisignal.2004088.

[17] Gates C, Ulintz P. 2016. Connor-Deduplication based on custom inline DNA barcodes. Available at https:// github.com/ umich-brcf-bioinf/Connor.

[18] Girardot C, Scholtalbers J, Sauer S, Su S-Y, Furlong EEM. 2016. Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. BMC Bioinformatics 17:419 DOI 10.1186/s12859-016-1284-2.

[19] Hadfield J. 2016. Increased read duplication on patterned flowcells- understanding the impact of exclusion amplification. Available at http:// core-genomics.blogspot.ru/ 2016/ 05/ increased-read-duplication-on-patterned.html (accessed on 7 January 2017).

[20] Hamilton JP, Potter JJ, Koganti L, Meltzer SJ, Mezey E. 2014. Effects of vitamin D3 stimulation of thioredoxin-interacting protein in hepatocellular carcinoma: TXNIP, vitamin D3 and HCC. Hepatology Research 44:1357-1366 DOI 10.1111/hepr.12302.

[21] Jang S-I, Lee Y-W, Cho C-K, Yoo H-S, Jang J-H. 2013. Identification of target genes involved in the antiproliferative effect of enzyme-modified ginseng extract in HepG2 hepatocarcinoma cell. Evidence-Based Complementary and Alternative Medicine 2013:1-8.

[22] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 14:R36 DOI 10.1186/gb-2013-14-4-r36.

[23] Kim H, Hwang J-S, Lee B, Hong J, Lee S. 2014. Newly identified cancer-associated role of human neuronal growth regulator 1 (NEGR1). Journal of Cancer 5:598-608 DOI 10.7150/jca.8052.

[24] Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, Taipale J. 2011. Counting absolute numbers of molecules using unique molecular identifiers. Nature Methods 9:72-74 DOI 10.1038/nmeth.1778.

[25] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078-2079 DOI 10.1093/bioinformatics/btp352.

[26] Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15:550 DOI 10.1186/s13059-014-0550-8.

[27] Morley S, You S, Pollan S, Choi J, Zhou B, Hager MH, Steadman K, Spinelli C, Rajen- dran K, Gertych A, Kim J, Adam RM, Yang W, Krishnan R, Knudsen BS, Di Vizio D, Freeman MR. 2015. Regulation of microtubule dynamics by DIAPH3 influences amoeboid tumor cell mechanics and sensitivity to taxanes. Scientific Reports 5:12136 DOI 10.1038/srep12136.

[28] Niu B, Fu L, Sun S, Li W. 2010. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics 11:187 DOI 10.1186/1471-2105-11-187.

[29] Roychowdhury S, Chinnaiyan AM. 2016. Translating cancer genomes and transcrip- tomes for precision oncology: translating genomics for precision oncology. CA: A Cancer Journal for Clinicians 66:75-88 DOI 10.3322/caac.21329.

[30] Saito T, Ichimura Y, Taguchi K, Suzuki T, Mizushima T, Takagi K, Hirose Y, Nagahashi M, Iso T, Fukutomi T, Ohishi M, Endo K, Uemura T, Nishito Y, Okuda S, Obata M, Kouno T, Imamura R, Tada Y, Obata R, Yasuda D, Takahashi K, Fujimura T, Pi J, Lee MS, Ueno T, Ohe T, Mashino T, Wakai T, Kojima H, Okabe T, Nagano T, Motohashi H, Waguri S, Soga T, Yamamoto M, Tanaka K, Komatsu M. 2016. p62/Sqstm1 promotes malignancy of HCV-positive hepatocellular carcinoma through Nrf2-dependent metabolic reprogramming. Nature Communications 7:12030 DOI 10.1038/ncomms12030.

[31] Shiroguchi K, Jia TZ, Sims PA, Xie XS. 2012. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proceedings of the National Academy of Sciences of the United States of America 109:1347-1352 DOI 10.1073/pnas.1118018109.

[32] Taniguchi K, Yamachika S, He F, Karin M. 2016. p62/SQSTM1-Dr. Jekyll and Mr. Hyde that prevents oxidative stress but promotes liver cancer. FEBS Letters 590:2375-2397 DOI 10.1002/1873-3468.12301.

[33] Tischler G, Leonard S. 2014. biobambam: tools for read pair collation based algorithms on BAM files. Source Code for Biology and Medicine 9:13 DOI 10.1186/1751-0473-9-13.

[34] Xu H, Luo X, Qian J, Pang X, Song J, Qian G, Chen J, Chen S. 2012. FastUniq: a fast de novo duplicates removal tool for paired short reads. PLOS ONE 7:e52249 DOI 10.1371/journal.pone.0052249.

[35] Zhang S, Liang X, Zheng X, Huang H, Chen X, Wu K, Wang B, Ma S. 2014. Glo1 genetic amplification as a potential therapeutic target in hepatocellular carcinoma. International Journal of Clinical and Experimental Pathology 7:2079-2090.

[36] Zhang T, Luo Y, Liu K, Pan L, Zhang B, Yu J, Hu S. 2011. BIGpre: a quality assessment package for next-generation sequencing data. Genomics Proteomics Bioinformatics 9:238-244 DOI 10.1016/S1672-0229(11)60027-2.

[37] Zucman-Rossi J, Benhamouche S, Godard C, Boyault S, Grimber G, Balabaud C, Cunha AS, Bioulac-Sage P, Perret C. 2007. Differential effects of inactivated Axin1 and activate β catenin mutations in human hepatocellular carcinomas. Oncogene 26:774-780 DOI 10.1038/sj.onc.1209824.