Mini Review
Open Access

Toward the integration of mate pair and RNA sequencing to identify gene fusions in cancer research: a mini review

Abstract

Mate pair (MPseq) and RNA sequencing (RNAseq) are important next-generation sequencing (NGS) techniques that are utilized to provide insight into tumorigenesis. Currently, MPseq is being successfully utilized in the clinic to predict chromosomal rearrangements while RNAseq is extensively used in the identification of gene expression, transcript expression and fusion detection. One of the strengths of MPseq is the fact that the fragments are longer than conventional pair-end fragments. This provides better coverage of genomic events such as structural variations. Fusions are structural rearrangements where there is an exchange of DNA sequences between genes. These kind of chromosomal rearrangements have great clinical importance. They are considered important biomarkers in neoplasia as well as therapeutic targets. However, as previously reported, fusion prediction tends to be difficult. This has been attributed to the large number of false positives due to sequencing errors. There are other factors such as poor alignment, library preparation, and insufficient depth of coverage. In addition, fusion predictions based purely on DNA technologies do not include gene expression information. Although, multiple software packages have been developed for fusions prediction, in many cases a consensus approach is required to eliminate false positives. MPseq's capabilities to detect genomic structural rearrangements can provide an unbiased orthogonal approach to predicting fusions when combined with RNAseq. In this mini-review we explore the benefits of MPseq and RNAseq as two complementary tools in the prediction of gene fusions.

Keywords: next-generation sequencing, mate pair, RNA sequencing, structural rearrangements, fusions, cancer

Volume 5 Issue 6 - 2017

Carlos P Sosa,[1,2] Daniel N Sosa,[3] George Vasmatzis[1,2,4]

[1]Biomarker Discovery Group, Center for Individualized Medicine, USA
[2]Bioinformatics and Computational Biology, University of Minnesota, USA
[3]CSAIL Massachusetts Institute of Technology, USA
[4]Department of Molecular Medicine, Mayo Clinic, USA

Correspondence: George Vasmatzis, Biomarker Discovery Group, Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota, 200 First Street SW, Rochester, Minnesota 55905, USA, Email Vasmatzis.George@mayo.edu

Received: June 05, 2017 | **Published:** June 15, 2017

Abbreviations

NGS, next-generation sequencing; WGS, whole genome sequencing; RNAseq, RNA sequencing; WES, whole exome sequencing; MP, mate pair; PE, pair-end; MPseq, mate pair sequencing; SNP, single nucleotide polymorphism; PTCL, peripheral t-cell lymphoma; IMTs, inflamatory myofibroblastic tumors

Introduction

Next-generation sequencing (NGS) has proven to be a powerful technique in advancing the field of cancer cytogenetics.[1] These significant advances in NGS have been coupled with the rapid decrease in the cost of the technology. This has opened the door for laboratories around the world to apply sequencing to a wide variety of diseases including cancer. NGS has allowed researchers to look at chromosomes and chromosome abnormalities even at the nucleotide level.[2] Although there are many variations that researchers study when it comes to alterations in the human genome, fusions play an important role in the study of cancer cells. Fusion discovery occurred more than 50years ago in the work of Nowell and Hungerford with patients that suffered from chronic myeloid leukemia.[3] Today NGS provides an opportunity to study fusions in a way that it was not possible just a few years ago. Researchers have available different types of sequencing technologies and they are applied depending on the problem at hand. Some of these technologies are: whole genome sequencing (WGS), whole exome sequencing (WES), and transcriptome sequencing, also called RNA sequencing (RNAseq).[4] WGS and WES have been applied mainly to determine point mutations, insertions and deletions, and copy number variants.[1] On the other hand, RNAseq focuses on abundances of all genes and transcripts in the cell.[4]

One of the libraries used in DNA sequencing technology that has been reported in the literature as an efficient, cost-effective method to determine structural variants is mate pair (MP).[5–7] One of the key differentiators between mate pair and other sequencing libraries is that in mate pair the DNA is fragmented into 2 to 5kilobases (Kb) segments.[8] Hereinafter, we refer to the insert as the segment of DNA or cDNA that encompasses the forward read (R1) and the reverse read (R2), and it is located between adaptors. The inner mate distance is the gap between R1 and R2.[9] Finally the fragment includes the adaptors.[9] Another more technical advantage is the fact that the insert in mate pair provides information on structural variants for longer regions than traditional pair-end inserts might not be able to detect.[8] In other words, longer fragments are more likely to provide deeper breakpoint coverage in certain genomic events. As previously reported, the chemistry involved in generating this type of library is different from a more commonly used library such as pair-end (PE) sequencing.[10] In mate pair the protocol involves a biotinylated enzyme to attach adapter to the ends of the DNA segment. The use of mate pair for doing NGS analysis is what is referred as mate pair sequencing (MPseq). The workflow is similar to other DNA workflows where initial reads are mapped to a genome reference.[5] In a post-mapping step structural variant detection, analysis, and visualization are performed.[6] This approach has been successfully applied to detect structural variations.[11]

RNAseq is another sequencing technique that has become very popular for quantifying RNA in a sample.[12] RNASeq has been applied to the prediction of alternative gene spliced transcripts, gene fusion, and single nucleotide polymorphism (SNP). However, RNAseq is extensively used to predict gene and transcript abundance for multiple conditions.[12] In the case of fusions RNAseq has proven to be a powerful technique and it has been successfully applied in cases such as *TMPRSS2-ERG*.[13] However, there is still opportunity for performance improvement. Kumar et al.,[14] have shown that when comparing 12 different software packages they observed performance dependency in quality, read lengths, and supporting reads. More recently, Haas et al.,[15] have compared 16 different packages to assess fusion prediction. In their work they used simulated as well as publicly available diverse cancer cell lines. One of the parameters they tested was read length. They looked at 50 base pairs and 101 base pairs in length. They found that longer read lengths improve performance. They also reported a consensus approach helped eliminate false positives. The threshold that they used was based on similar predictions by at least four of the different fusion predictors.

Discussion

MPseq and RNAseq have proven to be very successful techniques on their own for predicting structural variations and fusions, respectively. However, when the two techniques are combined they can leverage each other's strengths. It has been reported in the literature that a combined approach tends to overcome many of the limitations found when the methods are applied separately.[16] MPseq and RNAseq have successfully identified fusions in peripheral T-cell Lymphoma (PTCL).[17] One of the important points in this paper is that structural rearrangements predicted by MPseq do not necessarily involve genes nor expressed genes as fusion partners. On the other hand, RNAseq tends to identify expressed gene fusions.[17]

A study of Inflamatory myofibroblastic tumors (IMTs) is an example of complex structural rearrangements in the identification of the *TPM3-ALK* fusion.[18] In the combined approach MPseq is utilized first to identify all the structural rearrangements in the cancer sample. Its larger fragments provide better coverage for genomic events such as breakpoints. This is followed by further analysis to classify all the events. In the IMTs study a total of 209 events were identified and after further analysis 67% corresponded to intra-chromosomal events and 33% were classified as translocations.[18] In this case, although, the *ALK* gene showed several breakpoints, there was no event that would associate *ALK* with another gene. Further analysis showed that events affecting introns near *TPM3* correspond to the *ALK-TPM3* fusion. RNAseq analysis confirmed the existence of this fusion.[18]

Summary

MPseq is capable of detecting genomic events across the entire genome. Upon further analysis all these events can be classified an analyzed, including fusions. RNAseq provides information about expressed gene functions. Conversely, fusions predicted via RNAseq can be corroborated using MPseq.

Acknowledgements

None.

Conflict of interest

The author declares no conflict of interest.

References

1. Shyr D, Liu Q. Next generation sequencing in cancer research and clinical application. *Biol Proced Online*. 2013;15(4):1–11.
2. Gisselsson D. Chapter 2 Cytogenetic methods. In: Heim S, Mitelman F, editors. *Cancer Cytogenetics*. 4th ed. UK: John Wiley & Sons; 2015. p. 11–18.
3. Nowell PC, Hugenford DA. A minute chromosome in chronic granulocytic leukemia. *Science*. 1960;132(3438):1488–1501.
4. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. *J Appl Genet*. 2011;52(4):413–435.
5. Drucker TM, Johnson SH, Murphy SJ, et al. BIMA V3:an aligner customized for mate pair library sequencing. *Bioinformatics*. 2014;30(11):1627–1629.
6. Cradic KW, Murphy SJ, Drucker TM, et al. A simple method for gene phasing using mate pair sequencing. *BMC Med Genet*. 2014;15(19):1–8.
7. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next–generation sequencing. *Nat Methods*. 2009;6(11 Suppl):S13–S20.
8. http://www.illumina.com/documents/products/technotes/technote_nextera_matepair_data_processing.pdf.
9. http://thegenomefactory.blogspot.com/2013/08/paired–end–read–confusion–library.html
10. http://www.illumina.com
11. Graham RP, Barr Fritcher EG, Pestova E, et al. Fibroblast growth factor receptor 2 translocations in intrahepatic cholangiocarcinoma. *Hum Pathol*. 2014;45(8):1630–1638.
12. Griffith M, Walker JR, Spies NC, et al. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLoS Comput Biol*. 2015;11(8):1–20.
13. Maher CA, , Kumar–Sinha C, Cao X, et al. Transcriptome Sequencing to Detect gene Fusions in Cancer. *Nature*. 2009;458(7234):97–101.
14. Kumar S, Duy Vo A, Qin F, et al. Comparative assessment of methods for the fusion transcripts detection from RNA–Seq data. *Scientific Reports*. 2016;6(21597):1–10.
15. Haas BJ, Dobin A, Strnsky N, et al. STAR–Fusion: Fast and Accurae Fusion Transcript Detection from RNA–Seq. *bioRxiv*. 2017:120295.
16. Zhang J, White NM, Schmidt H, et al. INTEGRATES: Gene gene fusion discovery using whole genome and transcriptome data. *Genome Res*. 2015;11:1–29.
17. Boddicker R, Razidlo GL, Dasari S, et al. Integrated mate–pair and RNA sequencing identifies novel, targetable gene fusions in peripheral T–cell lymphoma. *Blood*. 2016;128(9):1234–1245.
18. Mansfield AS, Murphy SJ, Harris FR, et al. Chromoplectic TPM3–ALK rearrangement in a patient with inflammatory myofibroblastic tumor who responded to ceritinib after progression on crizotinib. *Ann Oncol*. 2016;11:2111–2117.