SMGroup

SM Journal of Bioinformatics and Proteomics

Editorial

A Lesson from Unmapped Reads in Next-Generation Sequencing Data

Deepak Singla

Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute PUSA, India

Article Information

Received date: May 20, 2016
Accepted date: May 21, 2016
Published date: May 23, 2016

***Corresponding author**

Deepak Singla, Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute PUSA, India, Email: deepkumar1983@gmail.com

Editorial

Next Generation Sequencing (NGS) technology is becoming very popular in generating the short reads genomic or transcriptome data. Based on the type of sequencing technology, it offers diverse applications varying from genome assembly, gene expression analysis to epigenetic changes [1-4]. Evolution of NGS technology enables us to provide deep insight to understand genomic variations, protein-nucleic acid interactions, meta-genome analysis etc.

Basically, the process starts with either de-novo or reference based assembly of sequencing reads and infers the relationship from the mapped reads such as SNPs, InDels, genome annotation [5,6]. Afterwards, a significant portion of reads were left as unmapped on reference genome that may be due to contamination or low quality bases. Previously, these unmapped reads were overlooked for their potential in genome analysis. Recently, in addition to mapped reads, these were also explored for their significant contribution in the genome of sequenced organism.

Analysis of the unmapped reads from 1000 genome project suggested that an average 0.13% of reads showed similarity to non-human genomes [7]. Similarly, Merchant *et al.* explored the unmapped reads from *B.tauras* genome and observed more than 150 contigs from microbial contamination [8]. Based on that, a new assembly version of *B.taurus* was deposited in NCBI database. In 2015, a large RNAseq dataset that comprised of 1873 cancer patient and 536 normal individual were analysed and ~2500 novel transcripts were found, majority of being long non-coding RNA (lnRNA) [9]. Likewise, Gouin *et al.* analyzed the re-sequencing data of 33 pea aphid genomes and established the relationship of unmapped reads of aphid genome with host plant specificity [10]. These analyses suggested the application of garbage data (unmapped reads) to excavate the meaningful information and therefore, demand the new software for the analysis of unmapped read data with high precision.

Previously, Ouma *et al.* used the SHRIMP software for alignment of unmapped reads because of its ability to handle highly polymorphic genomic region in the genome [11]. In the past, tools have been developed for the fast and accurate detection of microbial contamination in host genome [12,13]. Recently, Peng *et al.* developed open source software RAUR (Re-align the Unmapped Reads) for re-alignment of unmapped reads [14].

Conclusion

These studies entail the need to revisit the sequencing data for the identification of microbial contamination and novel transcripts that were not observed in previously assembled genome. The identified contigs from unmapped reads would also be used to fill the gap in genome assembly.

References

1. Liu H, Wang T, Wang J, Quan F, Zhang Y: Characterization of Liaoning cashmere goat transcriptome: sequencing, de novo assembly, functional annotation and comparative analysis. PLoS One. 2013; 8: e77062.
2. Lee J-H, Lee T, Lee H-K, Cho B-W, Shin D-H, Do K-T, et al. Thoroughbred Horse Single Nucleotide Polymorphism and Expression Database: HSDB. Asian-Australasian J Anim Sci. 2014; 27: 1236-1243.
3. Doherty R, Couldrey C. Exploring genome wide bisulfite sequencing for DNA methylation analysis in livestock: a technical assessment. Front Genet. 2014; 5: 126.
4. Jiang Y, Xie M, Chen W, Talbot R, Maddox JF, Faraut T, et al. The sheep genome illuminates biology of the rumen and lipid metabolism. Science. 2014; 344: 1168-1173.
5. Hai DT, Thanh ND, Trang PTM, Quang LS, Hang PTT, Cuong DC. Whole genome analysis of a Vietnamese trio. J Biosci. 2015; 40: 113-124.
6. Faber-Hammond JJ, Brown KH. Pseudo-De Novo Assembly and Analysis of Unmapped Genome Sequence Reads in Wild Zebrafish Reveal Novel Gene Content. Zebrafish. 2016; 13: 95-102.

7. Tae H, Karunasena E, Bavarva JH, McIver LJ, Garner HR. Large scale comparison of non-human sequences in human sequencing data. Genomics. 2014; 10: 453-458.

8. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. PeerJ. 2014; 2: e675.

9. Kazemian M, Ren M, Lin J-X, Liao W, Spolski R, Leonard WJ. Comprehensive assembly of novel transcripts from unmapped human RNA-Seq data and their association with cancer. Mol Syst Biol. 2015; 11: 826.

10. Gouin A, Legeai F, Nouhaud P, Whibley A, Simon J-C, Lemaitre C. Whole-genome re-sequencing of non-model organisms: lessons from unmapped reads. Heredity (Edinb). 2015; 114:494-501.

11. Rumble SM, Lacroute P, Dalca A V, Fiume M, Sidow A, Brudno M. SHRiMP accurate mapping of short color-space reads. PLoS Comput Bio. 2009; 5: e1000386.

12. Wood DE, Salzberg S. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014; 15: R46.

13. Kostic AD, Ojesina AI, Pedamallu CS, Jung J, Verhaak RGW, Getz G, et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol. 2011; 29: 393-396.

14. Peng X, Wang J, Zhang Z, Xiao Q, Li M, Pan Y. Re-alignment of the unmapped reads with base quality score.BMC Bioinformatics 2015;16: S8.