DEEP-LONG: A Fast and Accurate Aligner for Long RNA-Seq

Li Hou

Center for Bioinformatics

Yadong Wang (✉ ydwang@hit.edu.cn)

Harbin Institute of Technology, Harbin, China

Methodology

Keywords: Single molecule real time (SMRT), DEEP-LONG, RNA-Seq, sequencing technology

Posted Date: October 1st, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-79489/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

In recent years, because of the development of sequencing technology, long reads were widely used in many studies, include transcriptomics studies. Obviously, Long reads have more advantages than short reads. And long reads align also different from short reads align. Until now Lots of tools can process long RNA-Seq, but there still have some problems need to solve.

Results

We developed Deep-Long to process long RNA-Seq, Deep-Long is a fast and accurate tool. Deep-Long can handle troubles come from complicated gene structures and sequencing errors well, Deep-Long does well especially on [alternative splicing](#) and small exons. When sequencing error rate is low, Deep-Long can rapidly get more accurate results. While sequencing error rate rising, Deep-Long will use more time, but still more fast and accurate than most other tools.

Conclusions

Deep-Long is an useful tool to align long RNA-Seq to genome, and Deep-Long can find more exons and splices.

1 Introduction

Single molecule real time (SMRT) sequencing produced by Pacific Biosciences (PacBio) (1) and nanopore sequencing produced by Oxford Nanopore Technologies (ONT) (2) are the two main Third Generation Sequencing (TGS). Different from Second Generation Sequencing (SGS), TGS can directly obtain very long nucleotide sequences(1-100 kb) (3). Obviously, Long reads have more advantages than short reads. they can solve problems that used to be a challenge for earlier techniques (7). TGS can sequencing genomic regions such as high/low GC regions, tandem repeat regions, and interspersed repeat regions easily (4). Though the natural comparative advantages of long reads are very attractive, but the sequencing error of long reads much higher than SGS reads, High sequencing error rate makes accurate sequence analysis difficult. The error rate of raw PacBio data is 13–15% and 12–25% for ONT (5).

RNA-Seq(RNA sequencing) is an important approach to transcriptome profiling. SGS RNA-Seq help people to generate a global insight of the transcriptome and organization of many species. By using RNA-Seq, lots of novel transcribed regions were founded. RNA-Seq also help to find gene and exon boundaries, SNP(single-nucleotide polymorphism), RNA editing, gene fusion, gene expression and splicing isoforms of known genes (6). For complex transcriptomes, such as high/low GC regions and long sequence variations, it's hard for short reads to resolve them, while long reads can resolve these problems easier. Compere with short reads, although long reads produce low coverage and high sequencing error, but superior in identifying transcript isoforms, and gene expression quantification, and long reads is

especially suitable for the detection of novel splice sites. In recent years, long reads has been widely utilized in the analysis of the genomes of various organisms including prokaryotic and eukaryotic (7).

RNA-Seq align is a fundamental method of RNA sequence analysis. Because of Respective characteristic of long reads and short reads, RNA-Seq align methods have many differences. Currently, lots of long RNA-Seq aligner could handle third generation sequencing data. BLAT is an early align method. Gmap is widely used in many studies. compare with BBmap and STAR, Gmap have better performance, and use less time (9). GraphMap, Magic-BLAST, Minimap2 and deSALT are recently works, they all use different reference index from traditional hash-index to reduce seeding time. Minimap2 collect same minimizers between reference and reads, and then find optimal seed chain, performs DP-based global alignment between adjacent anchors in a chain. Minimap2 can map DNA or long mRNA sequences against a large reference database. optimized seeding and DP(Dynamic Programming) strategies make Minimap2 align reads in a very short time. deSALT use de Bruijn graph-based index align seeds on all reads, use these seeds align results to construct an optimal transcript, according to the optimal transcript give reads final align results. Because the probability of different reads have same sequencing error at same site is very low, construct an optimal transcript can help reads across high error region and get more credible align results, so deSALT performance well while reads have high sequencing error rate.

Deep-Long pay more attention to process hard align regions, such as small exons and exons with more SNPs and RNA editings. Deep-Long align reads in two granularities with two reference indexes. First granularity, Deep-Long use BWT(Burrows-Wheeler transform) index align long consecutive regions on reads. The most significant advantage of BWT index is save storage, so Deep-Long can run on a desktop computer easily. For hard align regions, Deep-Long use a 8 bp hash index of reference expect to find more seed clues of exons, and then give more credible align results. In most cases Deep-Long can rapidly get more accurate results, especially on alternative splicing and small exons. When sequencing error is very high, for single reads, it's hard to find long consecutive seeds, Deep-Long will use more time on hard regions, and similar sequence make 8 bp hash seeds give wrong guidances.

With the development of sequencing technology, various strategies were applied to reduce sequencing error rate, the accuracy of PacBio ROI reads > 98%, ONT use INC-Seq strategy obtain nanopore reads median accuracy > 97%, so more high accuracy long reads will be used in future RNA-Seq studies. When the sequencing error rate is very high, we don't suggest to use Deep-Long, deSALT will be a better choice.

2 Method

Deep-Long adopt seed-and-extend strategy like most aligners. Seed-and-extend strategy is an effective way on spliced RNA-Seq alignment, through combine seeds from different regions to find splice site. Seeding step always cost plenty of time, different align tools choose different ways to reduce seeding time.

Deep-Long have five major steps: 1) finding MEM seeds use BWT index. 2) finding extend seeds on reference. 3) finding optimal candidate chains. 4) finding seeds use hash index. 5) Checking exon and

splice sites.

1) Finding MEM seeds use BWT index.

Deep-Long use BWT index to find overlapped Maximal Exact Match (MEM) seeds from the beginning to end of each read. MEM concept was used in STAR, Mummer and MAUVE. Search MEM seeds only to the unmapped regions of the read sequentially makes the STAR algorithm extremely fast. Due to the high sequencing error of TGS reads, lots of MEM seeds may align to wrong regions on reference and finally miss the right align result of entire read. A little overlap between seeds can prevent this deviation with hardly no more time cost. After that Deep-Long will give a more micromesh deep align of unmapped regions between MEM seeds. Thanks for BWA, we can use BWT index to find MEM seeds easily.

for each read R , and given reference genome sequence G , from the first base of read R , find MEM seeds (i, j, l) , R_i is the i -th base of read, G_j is the j -th base of reference, l is the length of seed. Seed (i, j, l) is the seed from i -th base of read and j -th base of reference have Maximal Exact Match seed length l , R_{i+l} is different from G_{j+l} . from $i+l$, $i+l-4$ and $i+l+4$ base of read, find next MEM seeds, until the end of reads. DEEP-LONG accept MEM seeds length more than 16 bp, for less than 16 bp MEM seeds (i, j, l) , from $i+1$ base of read find new MEM seeds.

Sometimes, this seed finding strategy may cause bias, so DEEP-LONG use non-overlap 25 bp length seeds to correct this bias. For high sequencing error rate reads, DEEP-LONG suggest each 10 base use 15 bp length seeds.

Extend each seed from both front and back direction to make sure each seed (i, j, l) is the longest exact match, so R_{i-1} is different from G_{j-1} , and R_{i+l+1} is different from G_{j+l+1} .

2) Finding extend seeds on reference.

Because of SNP, RNA editing and sequencing error base, RNA-Seq is not continuous compare with reference. for each seed (i, j, l) , DEEP-LONG find Extend seeds from both tails. For example, seed1 (i_1, j_1, l_1) is an Extend seed of seed (i, j, l) from back tail, $i+l = < i_1 < i+l+10$ and $j+l-5 = < j_1 < j+l+15, l_1 \geq 4$. choose the best combination of Extend seeds join into MEM seeds set, and continue to find Extend seeds of seed1 from back tail recursively

Extend seeds help to across SNP, SNV(single-nucleotide variant), RNA editing and sequencing error base regions, and get more information about exon area, and then help to pick out correct seed chain and get correct alignment results.

3) Finding optimal candidate chains.

DEEP-LONG calculate all seeds score, the initiation score of one seed (i, j, l) is $l \times \text{match score}$. After finding seed chain, the score of one seed is the score of the best seed chain end with this seed. The score of seed S_i can be calculate from equation.

$$f(i) = \max\{f(j) + s(i) - \alpha(i, j) - \beta(i, j)\}, i > j > 0.$$

$f(i)$ is seed S_i final score, $f(j)$ is seed S_j final score, $s(i)$ is the initiation score of one seed S_i , $\alpha(i, j)$ is the overlap of S_i and S_j , $\beta(i, j)$ is the gap score between S_i and S_j .

DEEP-LONG consider chain score more than $(0.7 \times \text{read length} \times \text{match score})$ or best chain score more than $(0.3 \times \text{read length} \times \text{match score})$ to get final alignment results. If sequencing error is very high, it's hard to find enough seeds to support correct results, and then fail to align this read.

4) Finding seeds use hash index.

An 8 bp-kmer hash index of reference help Deep-Long find more small exons and exons with more SNPs and errors. DEEP-LONG consider the blank regions of reads on candidate seeds chains are hard align regions, DEEP-LONG find all 8 bp seeds of blank regions of candidate seeds chains. Finding all 8 bp seeds within a region is faster than finding all 8 bp on entire reference, so this step use less time than hash index aligner.

After finding 8 bp seeds, DEEP-LONG also find extend seeds of these seeds, find best candidate chain and use DP strategy to fill blank.

5) Checking exon and splice sites.

Most splice sites have splice signal GT-AG/CT-AC, DEEP-LONG will consider splice signal influence, a splice signal will add extra score to alignment result score. Long RNA-Seq always across more than 2 exons, GT-AG and CT-AC come from different chain of DNA, Check splice sites to make sure splice sites from same RNA-Seq have splice signals of same chain.

Because of high error rate, exons less than 40 bp will cause more wrong results. DEEP-LONG will check exons shorter than 40 bp, if remove short exon can increase results score, DEEP-LONG will abandon this exon. While input reads error rate is a little high, the tail exon of read is more likely a wrong exon, DEEP-LONG suggest cut off less than 25 bp tail exons.

3 Result

deSALT supply abundant data and compositive and synthetical evaluation program. We use deSALT simulation scripts to simulated 5 RNA-seq long read datasets of human(GRCh38, version 94) and 5 RNA-seq long read datasets of mouse (GRCm38, version 94) with depth of 30, respectively termed as "PacBio ROI reads", "PacBio subreads", "ONT 2D reads", "ONT 1D reads" and "NS-ONT reads", with different sequencing error and read length. Beside "NS-ONT reads" generated by NanoSim(21), others generated by PBSim(22), and all detail parameters follow deSALT.

DEEPL(Deep-Long) and other three most frequently used tools, deSALT, Minimap2 and GMAP were applied on 10 simulated datasets and 2 real datasets. All tools run without any gene annotations.

software version and Command lines are provided in Supplementary. We also follow deSALT use five metrics to describe sensitivity, accuracy and performance of the aligners.

Base%: the proportion of bases being correctly aligned to their ground truth positions.

Exon%: the proportion of the exons being correctly mapped.

Read80%: the proportion of Read80% reads, that $NT/NG > 80\%$ and $NT/NP > 80\%$, where NG is the number of ground truth exons within the read, NP is the number of exons predicted by the alignment and NT is the number of true positive exons.

Read100%: the proportion of Read100% reads, that $NT/NG = 100\%$ and $NT/NP = 100\%$.

All alingers run on a server with Intel Xeon E4280 CPU at 2.0GHZ and 1 Terabytes RAM, running Linux Ubuntu 16.04 in 8 threads.

From the test results on simulated datasets, DEEPL have good performance on all data sets. DEEPL get most Read80% and Read100%. Base% and Exon% always a little lower than deSALT, because high sequencing error have more influence to single read align results, its hard to confirm accurate exon boundary. For different error rate datasets, DEEPL maintain better performance. But for NS_ONT data sets DEEPL have an unexpected performance, Read80% and Read100% are very low, and GMAP have same tendency. Its more likely sequencing error of the real data used to simulate data is too high, that DEEPL can hardly find seed combinations higher than process threshold score value, and finally fail to give align results. But especially it has the ability to handle many difficult issues.

For reads contain small exons, DEEPL also have good performance, DEEPL use 8 bp-kmer hash index to find more information of read, 8 bp-kmer can easier find small exons, so DEEPL can easily process reads with small exons. Furthermore, for reads contain alternative splicing, DEEPL still have good performance, sometimes lower than deSALT. DEEPL done a lot on check whether exons credible, and DEEPL align each read independently, splice pattern of other reads hardly affect align result. So DEEPL can process alternative splicing like normal splicing, there is nothing special.

Moreover, the number of exons within a read have no effect to DEEPL align result, reads with various kinds numbers of exons nearly have same performance. DEEPL give align results of reads depend on seeds finding, if seeds confirm one exon exist, DEEPL will show it in align result. DEEPL search 8 bp hash seeds on reference intensively, try to find all exons on reads.

Overall, sequencing error is the major effect to DEEPL performance. Numbers of exon, length of exons and transcript structure have little effect to align results, DEEPL can process them easily. Sequencing error also affect speed of DEEPL. Speed of DEEPL always slower than deSALT and Minimap2, but faster than GMAP. But while sequencing error is very low, DEEPL can even use less time than deSALT and Minimap2, because of structural characteristics of BWT index, when sequencing error rate is low, DEEPL can find accurate and unique MEM seeds with BWT index very fast.

Here we also test DEEPL on two sets of real data of human SRR11638299 and SRR3476690, and two sets of real data of mouse SRR7345558 and SRR7345562. SRR11638299 and SRR3476690 are sequenced by PACBIO_SMRT (PacBio RS II) from different samples. SRR7345558 and SRR7345562 are Single cell isoform sequencing from whole mouse tissue RNA-seq from same sample, but SRR7345558 sequenced by OXFORD_NANOPORE (MinION), and SRR7345562 sequenced by PACBIO_SMRT (Sequel).

We evaluate real data align results use different program with deSALT, deSALT only consider the first align result of the same read, we prefer to consider the best align result of the same read. ALL align tools we have tested can align most reads well. On real data DEEPL also have a good performance. DEEPL always find out most transcripts and exons same with annotation. DEEPL find exon ability is unacted on exon length. DEEPL also find more small exons length < 20 bp. When exon length < 15 bp, GMAP always have good performance. Don't like simulated data, DEEPL align result on OXFORD_NANOPORE real data also have good performance as well as PACBIO_SMRT real data. But GMAP have an obviously low exon and transcript level on OXFORD_NANOPORE real data. Time cost on real data always have same tendency with simulated data, while GMAP use notable less time on SRR11638299.

We notice GMAP align results of some data sets have an unanticipated performance. We test each unexpected results for many time for all aligners, and try to remove personal and other factors, make sure the results credible, but we still don't know why this happen.

Table 1
Results on real datasets.

SRR11638299								
total trans(203675)total exons(654365)								
Found	Exon (0,15)	Exon (15,20)	Exon (20,30)	Exon (40,50)	Exon ALL	Trans ALL	Output ALL	Time
GMAP	1031	509	1923	3938	159525	41810	210931	15873
minimap2	975	490	1900	3934	159722	41214	171280	9578
DeSALT	982	493	1878	3930	158700	40769	106475	8727
DEEPL	980	492	1909	3973	163320	43165	255181	28192
SRR3476690								
total trans(203675)total exons(654365)								
Found	Exon (0,15)	Exon (15,20)	Exon (20,30)	Exon (40,50)	Exon ALL	Trans ALL	Output ALL	Time
GMAP	1292	686	2478	4525	192620	65822	863162	100044
minimap2	1277	674	2573	4709	193241	65132	302551	7085
DeSALT	1300	691	2613	4676	190952	64160	245146	6036
DEEPL	1309	693	2632	4844	222420	76146	1933482	35627
SRR7345562								
total trans(135144)total exons(487411)								
Found	Exon (0,15)	Exon (15,20)	Exon (20,30)	Exon (40,50)	Exon ALL	Trans ALL	Output ALL	Time
GMAP	1653	697	2521	4825	208825	63211	919888	121918
minimap2	1506	653	2453	4738	198590	59064	156310	7722
DeSALT	1520	678	2457	4739	197470	58625	261725	10826
DEEPL	1637	718	2612	5106	233049	70156	1916608	49743
SRR7345558								
total trans(135144)total exons(487411)								
Found	Exon (0,15)	Exon (15,20)	Exon (20,30)	Exon (40,50)	Exon ALL	Trans ALL	Output ALL	Time

SRR11638299								
total trans(203675)total exons(654365)								
GMAP	495	215	802	1576	66937	23781	316717	28708
minimap2	542	238	866	1806	74938	25232	320031	1250
DeSALT	559	252	891	1856	74935	26082	168375	1449
DEEPL	577	259	988	2103	101631	33763	615443	8597

4 Conclusion

Deep-Long is inclined to process complicated region on transcript, Deep-Long prefer to pay more time and effort on more meaningful genes structure, so sequencing error rate have a great influence of Deep-Long results. Deep-Long inherit the advantages of use BWT index to find MEM seeds, fast and use less space, but the disadvantage is also evident. When the sequencing error rate is extremely high, Deep-Long can hardly find seed combines for a single read, as this time deSALT is more efficient. While sequencing error rate of RNA-Seq is not very high, Deep-Long is fast and accurate tool. Deep-Long align each read independently, through 8 bp hash index, Deep-Long can get more information related to splicing and small exons, which makes Deep-Long more sensitivity.

BWT index help Deep-Long process low error rate region, and hash index help Deep-Long overcome hard region. So Deep-Long have tolerability of sequencing error. Although with sequencing error rates increase, the performance of Deep-Long have some reduce, and use more time, but Deep-Long still have strong ability on alternative splicing and small exons finding.

At present, lots of tools can use SGS reads to reduce long reads sequencing error rates, such as LCS. Along with the development of sequencing technology, more studies work on reducing sequencing error rates, like PacBio ROI reads and INC-Seq, and more low sequencing error rate reads will be used in future studies, Deep-Long will be more efficient on these data.

List Of Abbreviations

SMRT: Single molecule real time

PacBio : Pacific Biosciences

ONT : Oxford Nanopore Technologies

TGS : Third Generation Sequencing

SGS : Second Generation Sequencing

RNA-Seq: RNA sequencing

DP: Dynamic Programming

BWT : Burrows-Wheeler transform

MEM : Maximal Exact Match

SNP :single-nucleotide polymorphism

SNV : single-nucleotide variant

DEEPL : Deep-Long

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The source code of DEEPL are available at: <https://github.com/cathy-houli/DEEPL>.

The data simulation and benchmarking scripts and evaluation programs are available at: <https://github.com/hitbc/deSALT>.

The real data evaluation programs are available at: <https://github.com/cathy-houli/DEEPL>.

Command and more statistical result are available in supplementary file.

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been supported by the National Key Research and Development Program of China (Nos: 2017YFC1201201 and 2017YFC0907503).

Authors' contributions

LH designed the method, implemented the method and performed the analysis. LH and YW wrote the manuscript.

Acknowledgements

We appreciate the advice and discussion from Professor Liu Bo. The BWT indexing and BWT array search procedure of Deep-Long were modified from BWA. Thanks deSALT for providing simulated data and evaluation scripts.

References

1. Rhoads A, Au KF: PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* 2015, 13:278-289.
2. Mikheyev AS, Tin MM: A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* 2014, 14:1097-1102.
3. Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang XJ, Buck D, Au KF: Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res* 2017, 6:100
4. Nakano K, Shiroma A, Shimoji M, et al. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area[J]. *Human Cell*, 2017, 30(3):149-161.
5. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ: Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol Detect Quantif* 2015, 3:1-8.
6. Wang Z, Gerstein M, Snyder M: RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009, 10:57-63.
7. Boldogkői, Zsolt, Moldován, Norbert, Balázs, Zsolt... : Long-Read Sequencing – A Powerful Tool in Viral Transcriptome Research. *ScienceDirect*, 2019, 578-592.
8. Douglas R. Hurst, Alka Mehta, Yi Xie, et al. Identification of differentially expressed Breast Cancer Metastasis Suppressor 1 (BRMS1) splice variants in cell strains from MCF10A[J]. *cancer research*, 2006, 66:1012-1012.
9. Križanovic K, Echchiki A, Roux J, Šikic M: Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics* 2018, 34:748-754
10. Kent, & W., J. . (2002). Blat—the blast-like alignment tool. *Genome Research*, 12(4), 656-664.
11. Wu TD, Watanabe CKJB: GMAP: a genomic mapping and alignment program for mRNA and EST sequences. 2005, 21:1859-1875.
12. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013, 29:15-21.

13. Bushnell B: BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States); 2014.
14. Maric J, Sovic I, Krizanovic K, Nagarajan N, Sikic MJb: Graphmap2-splice-aware RNA-seq mapper for long reads. 2019:720458.
15. Grzegorz M Boratyn, Jean Thierry-Mieg, Danielle Thierry-Mieg, Ben Busby, Thomas L Madden:Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads. BMC Bioinformatics. 2018.
16. Li HJB: Minimap2: pairwise alignment for nucleotide sequences. 2018, 34:3094-3100.
17. Liu, Bo and Liu, Yadong and Zang, Tianyi and Wang, Yadong. deSALT: fast and accurate long transcriptomic read alignment with de Bruijn graph-based index. BMC Bioinformatics. 2019.
18. Delcher, A.L., et al. Fast algorithms for large-scale genome alignment and comparison, Nucleic Acids Res, 2002, 30, 2478-2483.
19. Darling, A.C., et al. Mauve: multiple alignment of conserved genomic sequence with rearrangements, Genome Research, 2004, 14, 1394-1403.
20. Li C , Chng K R , Boey E J H , et al. INC-Seq: accurate single molecule reads using nanopore sequencing[J]. GigaScience, 2016, 5(1):1-11.
21. Yang C, Chu J, Warren RL, Birol IJG: NanoSim: nanopore sequence read simulator based on statistical characterization. 2017, 6:gix010.
22. Ono Y, Asai K, Hamada MJB: PBSIM: PacBio reads simulator—toward accurate genome assembly. 2012, 29:119-121.

Figures

Figure 1: MEM seeds and Extend seed.

Figure 1

Figure 1

Reference(correct transcript)

Figure 2: Check exons. A) part of blue exon and part of green exon make up a wrong exon align to wrong region. B) part of green exon align to wrong region.

Figure 2

Figure 2

Figure 3: Results on simulated datasets

Figure 3

Figure 3

Figure 4: Time cost on simulated datasets

Figure 4

Figure 4

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [DEEPLSupplementaryInformation.docx](#)