INNOVARE ACADEMIC SCIENCES
Knowledge to Innovation

International Journal of Pharmacy and Pharmaceutical Sciences

ISSN- 0975-1491 Vol 8, Issue 2, 2016

Original Article

A *DE NOVO* ASSEMBLY METHOD FOR SHORT SEQUENCE OF SOLID-SAGE READS RESPONSIBLE FOR WHEAT (*TRITICUM AESTIVUM* L.) LEAF RUST

JYOTI PATHAK, KUNAL MUKHOPADHYAY, RAJU PODDAR*

Department of Bio-Engineering, Birla Institute of Technology, Mesra, Ranchi 835215, Jharkhand, India
Email: rpoddar@bitmesra.ac.in

Received: 30 Oct 2015 Revised and Accepted: 30 Dec 2015

ABSTRACT

Objective: Wheat leaf rust is one of the most widespread rust diseases caused by *Puccinia triticina* Eriks. *De novo* assembly of short sequence reads in order to understand the molecular phenomenon underlying wheat leaf rust interaction and to assemble differentially expressed genes, resistance genes and the genes encoding transcription factors in response to *Puccinia* infection in wheat was the main objective of the present study.

Methods: *De novo* assembly of SOLiD (sequencing by oligonucleotide ligation and detection) SAGE (serial analysis of gene expression) sequence reads from a pair of Near-isogenic lines (NILs) of wheat cultivar HD2329 with *Lr*28 (resistant) and HD2329 lacking *Lr*28 (susceptible) that were either infected with the most virulent pathogen *Puccinia triticina* or inoculated as mock in the absence of any reference sequence was carried out using multiple k-mer approach. Combinations of different software working on different algorithm were used to obtain a maximum number of differentially expressed transcripts.

Results: *De novo* assembly at different k-mers produced a large number of contigs. The size of contigs was further increased with the use of different assembly software. Redundancy was removed both at nucleotide and protein levels, which increased the quality of assembly.

Conclusion: For the assembly of short sequences of the complex genome such as those of polyploids a combination of software gives longer and unique contigs. It may be used in understanding the molecular mechanism of plant-microbe interaction.

Keywords: Wheat, Leaf rust, SOLiD, SAGE, *De novo* assembly, NILs.

INTRODUCTION

Wheat is one of the major cereals for human nutrition. Wheat consumption worldwide is estimated to surpass 817 million tons by 2030 [1]. Rust diseases are the most widespread and economically important diseases of cereal crops worldwide. Three distinct diseases, leaf rust (Brown rust), stripe rust (Yellow rust) and stem rust (Black rust), and found in wheat. The fungal pathogens that cause these diseases are constantly evolving new virulent form of rust pathotypes and overcome the resistance of wheat varieties. Leaf rust caused by the fungus *Puccinia triticina Erikss. & Henn.* is the most common and widely distributed of the three wheat rusts. For the control of a mutagenic variety of leaf rust pathogens, more knowledge will be required on the wheat sequences and their function such as in plant-pathogen interaction, signaling mechanism and role of transcription factors in obtaining a better combination of *Lr* genes and enhanced defense activity during infection.

Wheat genome size is about 17 GB. Whole genome sequence data are still not available. Transcriptome sequencing is considered to be an effective alternative for rapid identification of wheat genes. In the absence of complete reference sequence, *De novo* assembly of sequencing reads is a key step for comparative study.

De novo sequence assembly is the method of generating contiguous sequences by merging together individual sequence reads. These contiguous sequences or contigs share the same nucleotide sequence as the template (DNA or RNA) from which they are derived. As next generation sequencing technologies are revolutionizing, there is a limitation to the assembly programs used for short sequence data of 20-30 nucleotides such as SOLiD reads. Several algorithms are reviewed for *De novo* assembly [2-5] and have been implemented by many software packages including CAP3 [6], Velvet [7], Oases [8], Trinity [9], Trans-ABySS [10], but none of them can be individually used for the assembly of solid reads. The main reasons are a large number of sequences generated which makes the analysis computationally costly and base calling errors are more likely to occur due to their short length. Wheat is an allohexaploid containing homeologous and duplicated genes. Wheat contains 80% of repetitive sequences. It becomes difficult to resolve repetitive sequences since; the reads generated by sequencing are shorter than the repetitive unit. This can lead to misassembled or partial assembled contigs. Another limitation to the assembly of short sequence reads is that most of the assemblies have been reported with long sequences [11] or with a combination of short and long reads [12]. Short read technology such as Illumina has been successful in a wide variety of transcriptome analysis in plants with paired-end sequences [13, 14]. Assembly of single-ended solid sequences is difficult. However, studies [15, 16] support assembly of short sequence reads of 20-30 nucleotides long with gaps. Combinations of *De novo* assembler have been used to reconstruct high-quality transcriptome in polyploids [17]. We present a method for assembly of short reads where multiple k-mer and combination of assembly software's based on different algorithms is used. The basic assumption of this is that different k-mers will allow the assembly of transcripts with different abundances. At higher k-mer length more number of contig sequences are assembled which are highly expressed transcripts. On the contrary, at lower k-mer length poorly expressed transcripts are assembled better [7, 18].

MATERIALS AND METHODS

SAGE libraries were prepared for four selected wheat lines; susceptible HD2329 mock (S-M), susceptible HD2329 infected (S-PI), resistant HD2329+*Lr28* mock (R-M) and resistant HD2329+*Lr28* infected (R-PI) using a Solid SAGE kit (Applied Biosystems, CA, USA). These libraries are named SAGE1, SAGE2, SAGE3 and SAGE4 respectively.

Pre-processing of reads

The reads from the four libraries were imported separately to CLC Genomics Workbench v. 6.5.1. Trimming was performed under the following trim settings: (i) removal of the low-quality sequence (limit = 0.05), trimming of sequences on the basis of quality score on a phred scale. The quality score is converted into error probability. This has high value for low-quality bases. (ii) Removal of ambiguous nt: maximum of 2 nt allowed, (iii) removal of adapter sequences, (iv) removal of sequences on length: minimum length of 20 nt. Since two libraries S-PI and R-PI might contain reads originating from *Puccinia* transcripts, all post-trimmed reads were mapped to *Puccinia* transcripts available at The Broad Institute (https://www.broadinstitute.org/annotation/genome/puccinia_group/MultiHome.html). The reads that did not match to *Puccinia* transcripts were used for *de novo* assembly. Individual assembly of the four libraries was carried out on a server with 64 GB RAM. Fig. 1 represents the overall strategy for *de novo* assembly.

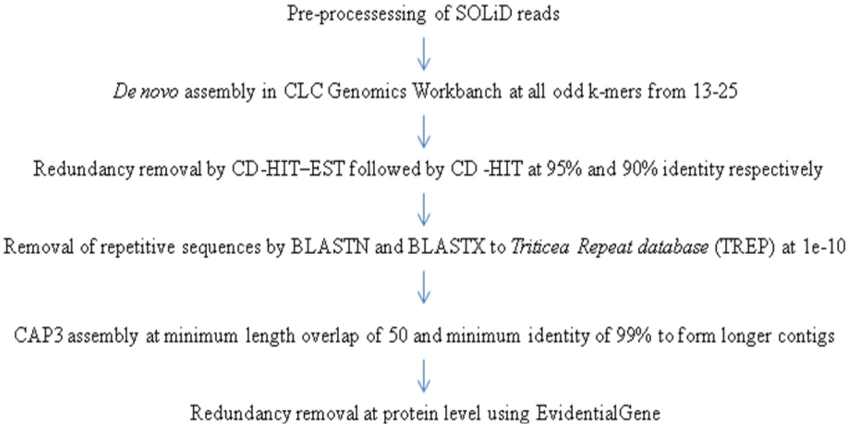
Pre-processessing of SOLiD reads
De novo assembly in CLC Genomics Workbanch at all odd k-mers from 13-25
Redundancy removal by CD-HIT–EST followed by CD -HIT at 95% and 90% identity respectively
Removal of repetitive sequences by BLASTN and BLASTX to *Triticea Repeat database* (TREP) at 1e-10
CAP3 assembly at minimum length overlap of 50 and minimum identity of 99% to form longer contigs
Redundancy removal at protein level using EvidentialGene

Fig. 1: Overview of methodology used for *De novo* assembly

De novo assembly

De novo assembly was carried out using different software's based on two different algorithms. CLC Genomics Workbench works on de Bruijn graph (DBG) and CAP3 working on an overlap consensus method (OLC). The OLC algorithm is computationally more intensive than DBG, and most effective for low coverage long read, whereas DBG is more suitable for high coverage short reads especially for large genome assembly. The use of k-mers which are shorter than the read lengths in de Bruijn graphs reduces the computational intensity of this method [19].

Reads from four libraries were taken, and *de novo* assembly was carried out at different odd k-mers from 13-25 in CLC Genomics Workbench v.6.5.1. CLC bio's *de novo* assembly algorithm works by using de Bruijn graph [7]. Reads were mapped back to the contigs with a minimum length fraction of 80% and minimum identity of 90% in order to update the contigs. The advantage of mapping reads to the contigs and selecting 'Update Contig' in CLC is that it removes the region in the contig where no read is mapped and for any mismatch to the contig, the sequence is updated to reflect the majority base at that location among the reads mapped there. Assembled sequences obtained from all k-mer were combined to get a maximum number of sequence assemblies. Redundant sequences were removed by using CD-HIT-EST followed by CD-HIT [20] at 95% and 90% identity respectively. Sequences greater than 100 base pairs were selected and taken for further process. Wheat genome is composed of 80% of repetitive sequences. Candidate repetitive elements and transposons were identified based on results from BLASTN and BLASTX searches against the nucleotide and protein *Triticea Repeat Sequence Databases* (TREP) using an E-value cutoff of 1e-10.

To produce longer, more complete contigs CAP3 was used for further assembly. Different overlap and identity cutoff were tried for assembly and then compared for similarity by BLASTN to wheat Unigene at an E-value cutoff 10^{-5}. Based on the results we obtained the best assembly at the minimum overlap of 50 base pairs and minimum identity of 99% with the following options "-p 99–l 50 v 100". Removal of redundant sequences at protein level was performed using Evidential Gene tr2aacds pipeline [21] (http://arthropods. eugenes.org/Evidential Gene/about/ Evidential Gene_assembly_pipe.html). It selects the 'best' set of *De novo* assembled transcripts, based on coding potential. The algorithm first produces CDS and amino acid sequences for each sequence and then removes redundant sequences using the amino acid information for choosing the best coding sequences amongst identical sequences. Self-on-self BLAST is then implemented to identify highly similar sequences. The alignment data and CDS/protein identities are then used to select and output transcripts classified as 'main' or 'alternate', and another set classified as 'dropped' which did not pass the internal filters. The primary and alternate *De novo* assembled transcripts were used for further assessments.

The assembly produced can be used as a reference for applications such as single nucleotide polymorphism (SNP) detection, expression study, gene identification and pathway study.

RESULTS

Table 1, 2, 3 and 4 provides the summary of the results obtained after *de novo* assembly at different k-mers all odd numbers starting from 13-25. In all four libraries, the largest number of contigs obtained at k-mer 13. However, the average length of contigs and N50 was greatest at k-mer 15. A contig N50 is calculated by first ordering every contig by a length from longest to shortest. Next, starting from the longest contig, the lengths of each contig are summed, until this running sum equals one-half of the total length of all contigs in the assembly. The contig N50 of the assembly is the length of the shortest contig in this list [22].

Table 1: Summary of *De novo* assembly in CLC genomics workbench at different k-mer for SAGE1 library

K-mer	Total contig	Avg. contig size	%matched reads	No. of transcript>1000bp	Min contig length	Max contig length	N50
13	1 70 576	879	45.57	59 358	16	8026	1299
15	27 336	1266	32.97	14 652	16	9394	1813
17	4331	594	16.55	824	17	16 203	1126
19	3182	463	14.48	384	18	12 858	915
21	5608	424	18.36	592	25	12 858	869
23	6310	389	25.58	615	26	14 238	844
25	3850	373	19.72	329	30	13 080	786
Total	2 21 193						

Table 2: Summary of *De novo* assembly in CLC genomics workbench at different k-mer for SAGE2 library

K-mer	Total contig	Avg. contig size	%matched reads	No. of transcript>1000bp	Min contig length	Max contig length	N50
13	1 67 587	886	47.66	58 870	16	9001	1307
15	23 483	1290	32.89	12 686	16	11 239	1853
17	4892	555	17.36	813	17	11 915	1167
19	3171	395	11.64	292	23	12 665	876
21	5324	367	17.27	429	25	13 783	817
23	5463	334	18.27	393	28	13 695	763
25	3319	317	20.47	210	28	8967	708
Total	2 13 239						

Table 3: Summary of *De novo* assembly in CLC genomics workbench at different k-mer for SAGE3 library

K-mer	Total contig	Avg. contig size	%matched reads	No. of transcript>1000bp	Min contig length	Max contig length	N50
13	1 61 161	888	48.87	56 577	16	8246	1302
15	12 362	1307	24.44	6729	16	10 817	1917
17	3072	500	11.97	454	20	11 348	1003
19	2560	401	9.49	240	25	6867	805
21	4209	351	14.17	315	25	9275	740
23	3923	330	16.16	252	23	14 238	686
25	2579	322	12.12	155	32	9338	658
Total	1 89 866						

Table 4: Summary of *De novo* assembly in CLC Genomics Workbench at different k-mer for SAGE4 library

K-mer	Total contig	Avg. contig size	%matched reads	No. of transcript>1000bp	Min contig length	Max contig length	N50
13	1 51 627	915	51.00	55 661	16	8504	1331
15	11 057	1251	24.04	5706	16	12 768	1873
17	2418	438	9.31	312	17	6824	1120
19	2160	325	9.35	172	24	6824	804
21	3311	295	13.16	202	25	8926	763
23	2705	265	13.99	140	32	6824	624
25	1481	237	10.04	54	35	4220	452
Total	1 74 758						

The total number of contigs reduced to about 57% after redundancy removal by CD-HIT (table 5). 1 06 816, 1 04 054, 92 625 and 88 309 numbers of contigs greater than 100 base pairs were selected in the four libraries respectively. 661, 636, 480 and 500 repetitive sequences were found after BLASTN and BLASTX to TREP. The sequences obtained after removal of repetitive sequences were assembled using CAP3 which formed 10, 17, 18 and 12 supercontigs in SAGE1, 2, 3 and 4 libraries respectively. Finally, after redundancy removal by Evidential Gene at the protein level, we obtained 45 263, 44 467, 41 556 and 40 023 final contigs in SAGE1, SAGE2, SAGE3 and SAGE4, respectively.

Table 5: Number of contigs after assembly and redundancy removal

Contigs	SAGE1	SAGE2	SAGE3	SAGE4
After CLC	2 21 193	2 13 239	1 89 866	1 74 758
After CD-HIT	1 25 520	1 22 023	1 07 943	1 01 470
>100 base pair	1 06 816	1 04 054	92 625	88 309
After TREP	1 06 165	1 03 418	92 145	87 809
After CAP3	1 06 155	1 03 401	92 127	87 797
After evidential gene	45 263	44 467	41 556	40 023

DISCUSSION

For non-model plant such as wheat lacking, well-defined genome de novo assembly is required for downstream analysis. We used two different algorithms for the assembly. DBG implemented in CLC which is suitable for high coverage short sequence reads and OLC implemented in CAP3 which works well on low coverage longer reads. To reduce the number of *de novo* assembled transcripts, CAP3 and Evi pipelines were used. *De novo* assembly may generate chimeric sequences for allopolyploid plant Transcriptome due to the presence of duplicated and homoeologous gene copies and transcripts, this assembly work gets further complicated due to the presence of transcript isoforms. Therefore, discrimination between homologous and paralogous copies of gene transcripts is important. Here, we have implemented a new protocol to overcome this difficulty of distinguishing homologous transcripts by varying the k-mer range and indirectly the effect of expression level compared to other methods [23-26]. The basis behind this protocol using Evi pipeline (which focuses on the coding potential rather than just the transcript length) is to get optimal transcripts. It will first generate as many *de novo* assembled transcripts as possible from a broad range of k-mers and then selecting from this a 'best' set of putative transcripts based on CDS and protein length. One of the disadvantages of using a multiple k-mer approaches is the high redundancy generated by duplicated genes and different A/B chimeric forms of the same gene assembled at different k-mer sizes. A large proportion of redundancy has been shown to be eliminated by CD-HIT. There is a greater reduction (about 55 %) of *de novo* assembled transcripts after using Evi pipeline as is evident from earlier studies [17] in an allotetraploid.

The N50 value and the total number of transcripts decreased with increasing k-mer ranges 15 to 19 suggesting over-representation at lower k-mer and under-representation at higher k-mers. It will also reduce the percentage of reads utilized in the assembly. This reported protocol is advantageous in terms of optimal k-mer findings (table 4 and 5) over earlier studies [25, 26] which reported that *de novo* transcriptome assembly is more fractionated in polyploid species.

CONCLUSION

We present a method for the assembly of single-ended Solid reads. Different assembly software were used for this purpose. Post processing of Solid-SAGE NGS reads and *De novo* assembly of four Solid SAGE libraries were successfully completed to produce a total of nearly 1,80,000 contigs. For the assembly of sequences derived from complex polyploids, a combination of assembly algorithms and software with assembler-specific optimal parameters can be used for the assembly of short sequence reads to obtain optimum results.

ACKNOWLEDGEMENT

We are thankful to the Department of Biotechnology, Govt. of India, (BT/BI/04/065/04 and BT/PR3925/BID/7/384/2011) for financial support.

CONFLICT OF INTERESTS

Declared none

REFERENCES

1. Zhang W, Liu G, Bai C. A forecast analysis on global production of staple crops; 2007.
2. Flicek P, Birney E. Sense from sequence reads methods for alignment and assembly. Nat Methods 2010;*7*:479.
3. Jackman SD, Birol I. Assembling genomes using short-read sequencing technology. Genome Biol 2010;11:202.
4. Pop M. Genome assembly reborn: recent computational challenges. Briefings Bioinf 2009;10:354-66.
5. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet 2008;24:142-9.
6. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res 1999;9:868-77.
7. Zerbino DR, Birney E. Velvet: algorithms for *De novo* short read assembly using de bruijn graphs. Genome Res 2008;18:821-9.
8. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust *De novo* RNA-seq assembly across the dynamic range of expression levels*.* Bioinformatics 2012;28:1086-92.
9. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 2011;29:644-52.
10. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, *et al. De novo* assembly and analysis of RNA-seq data. Nat Methods 2010;7:909-12.
11. Brenchly R, Spannagi M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, *et al.* Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 2012;491:705-10.
12. Reinhardt JA, Baltrus DA, Nishimura MT, Jeck WR, Jones CD, Dangl JL. *De novo* assembly using low-coverage short read sequence data from the rice pathogen pseudomonas syringae pv. Oryzae Genome Res 2009;19:294-305.
13. Duan J, Xia C, Zhao G, Jia J, Kong X. Optimizing *De novo* common wheat transcriptome assembly using short-read RNA-seq data. BMC Genomics 2012;13:392.
14. Gahlan P, Singh HR, Shankar R, Sharma N, Kumari A, Chawla V, *et al*. *De novo* sequencing and characterization of *Picrorhiza kurrooa* transcriptome at two temperatures showed major transcriptome adjustments. BMC Genomics 2012;13:126.
15. Whiteford N, Haslam N, Weber G, Prugel-bennett A, Essex JW, Roach PL, *et al.* An analysis of the feasibility of short read sequencing. Nucleic Acids Res 2005;33:e171-e171.
16. Chaisson M, Pevzner P, Tang H. Fragment assembly with short reads. Bioinformatics 2004;20:2067-74.
17. Nakasugi K, Crowhurst R, Bally J, Waterhouse P. Combining transcriptome assemblies from multiple *De novo* assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. PloS One 2014;9:e91776.
18. Gibbons JG, Janson EM, Hittinger CT, Johnston M, Abbot P, Rokas A. Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. Mol Biol Evol 2009;26:2731-44.
19. Li Z, Chen Y, Mu D, Yuan J, Shi Y, Zhang H, *et al.* Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. Briefings Funct Genomics 2012;11:25-37.
20. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658-9.
21. Gilbert D. Gene-omes built from mRNA seq not genome DNA. 7th annual arthropod genomics symposium. Notre Dame; 2013.
22. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 2012;13:329-42.
23. Gruenheit N, Deusch O, Esser C, Becker M, Voelckel C, Lockhert P. Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. BMC Genomics 2012;13:92.
24. Duan J, Xia C, Zhao G, Jia J, Kong X. Optimizing *de novo* common wheat transcriptome assembly using short-read RNA-Seq data. BMC Genomics 2012;13:392.
25. Krasileva K, Buffalo V, Bailey P, Pearce S, Ayling S, Tabbita F, *et al.* Separating homeologs by phasing in the tetraploid wheat transcriptome. Genome Biol 2013;14:R66.
26. Ranwez V, Holtz Y, Sarah G, Ardisson M, Santoni S, Glémin S, *et al.* Disentangling homologous contigs in allotetraploid assembly: application to durum wheat. BMC Bioinformatics 2013;14(Suppl 15):S15.