Plant Omics Journal POJ

POJ 8(6):581-586 (2015) ISSN:1836-3644

HCRA: a hybrid colour-space read-additive method for *de novo* transcriptomic assembly integrating Illumina and SOLiD datasets

Adriano Viegas[1], Artur Silva[2], Sylvain Darnet[1]*

[1]Federal University of Para, Institute of Biological Sciences, Laboratory of Plant Biotechnology, Belem, Para, Brazil
[2]Federal University of Para, Institute of Biological Sciences, Laboratory of DNA Polymorphism, Belem, Para, Brazil

Corresponding author: shd@laposte.net

Abstract

Hybrid approaches have been developed to mix large datasets from different next-generation sequencing (NGS) platforms, such as Illumina and SOLiD, for optimizing *de novo* transcriptomic assembly. The classical approach (CA) is to form supercontigs from contigs obtained by the *de novo* assembly of each dataset. We have developed a new hybrid colour-space read-additive (HCRA) approach to assemble both datasets encoded in colour-space reads, using a multiple k-mer Velvet/Oases method. All reads are combined in colour-space and subjected together to the assembler. To evaluate both CA and HCRA methods, we used assembly statistics such as total base pairs, N50, reads mapped back to transcripts, and percentage of unique transcripts partially and completely identified. These approaches were tested using SOLiD and Illumina simulated runs, generated from 41392 sequences of *Arabidopsis* cDNA, totalling 64.8 Mb and with an N50 of 1913 bp. The CA and HCRA methods generated contig datasets with 225811 and 172835 transcripts with N50s of 931 and 1617 bp, respectively. Compared with the initial *Arabidopsis* dataset, 35960 contigs were reconstructed with CA, totalling 35.4 Mb, and 35240 contigs were reconstructed with HCRA, totalling 52.3 Mb. The HCRA method generated approximately 2-fold longer contigs than CA, and 40% more transcripts were completely identified. The proposed pipeline was applied to a real dataset of the *Piper nigrum* transcriptome, generating 60645 unigenes with an N50 of 1653 bp and representing about 71 Mb of its transcriptome. This method improves the integration of SOLiD datasets with those from other NGS platforms and should open new perspectives to add colour-space datasets to Illumina runs to improve *de novo* transcriptomic assembly in non-model organisms.

Keywords: Next-generation sequencing, transcriptome, hybrid *de novo* assembly, *Arabidopsis thaliana*, *Piper nigrum*.
Abbreviations: EST_expressed sequence tag; HCRA_hybrid colour-space read-additive; CA_contig additive; CEGMA_core eukaryotic gene mapping approach; RMBT_reads mapped back to transcripts.

Introduction

RNA-seq is a powerful technology based on high-throughput next-generation sequencing (NGS) methods that describes expressed RNAs highly efficiently in an exhaustive matter (Wang et al., 2009). This technology is commonly used in plant genetics to obtain sequences of coding and non-coding RNAs, to estimate relative RNA expression, and to identify single nucleotide polymorphisms (Wang et al., 2009; Martin and Wang, 2011). A better and easier approach to obtain the complete transcriptomic sequence dataset using NGS short reads is to compare and align reads against genomes and EST references to facilitate read assembly (Garg and Jain, 2013), but RNA-seq has opened new perspectives for characterizing non-model plants. *De novo* sequencing, which sequences without previous knowledge of the genome and transcriptome, is useful and efficient for agrigenomic approaches and a rapid way to obtain new large genetic datasets of non-model plants. The limiting step of *de novo* sequencing is the bioinformatic analysis. *De novo* assembly of short reads is a complex and time-consuming challenge for plant data due to the large size of plant genomes, high number of paralogs and isoforms, polyploidy, and other genomic duplications (Garg and Jain, 2013). Many bioinformatic tools have been developed to detect, identify, and reconstruct RNA sequences with high accuracy and precision for optimizing *de novo* assembly. The more efficient tools for short reads are based on de Bruijin graphs, such as Velvet (Zerbino and Birney, 2008), Oases (Schulz et al., 2012), ABySS (Simpson et al., 2009), Trinity (Grabherr et al., 2011), CLC Genomics Workbench (www.clcbio.com), and SOAPdenovo-Trans (Xie, 2014). All *de novo* assembly tools are efficient, but the results, the predicted assembled transcriptomes, vary depending on the assembler tool and its parameters, illustrating that this bioinformatic step remains a limiting step of RNA-seq to evaluate plant transcriptomes (Wang et al., 2009; McGettigan, 2013). One way to improve the prediction of transcriptomic sequences is to test and combine the results of different transcriptomic assemblers or different parameters in assembly steps. The commonest method is to form supercontigs from contigs obtained by different assemblers from the same reads of the datasets. Transcriptomic prediction is based on two steps: the primary assembly to obtain the contigs from reads, and the secondary assembly to obtain the supercontigs by processing contigs with assembler tools (Wang et al., 2012). This method has several limitations, such as the formation of chimeric sequences and the lack of isoform-detection sensitivity. The alternative method is the contig additive (CA) approach, which is based on a clustering and not an assembler tool. Its principal objective is to remove sequence redundancy while preserving isoform prediction. For example, Surget-Groba and Montoya-Burgos (2010) have developed a strategy to join contigs obtained with multiple k-mer parameters from Velvet tools. An alternative, the hybrid approach, is to integrate datasets from different NGS platforms, exploiting the advantage of each platform relative to the other, such as sequencing chemistry and library preparation. The most common method for non-model plants is the integration of large datasets of short reads, such as SOLiD and Illumina datasets, with longer reads, such as 454 sequencer or Sanger data. Large short-read datasets with high sequencing quality are used to correct and extend long reads or contigs (Koren et al., 2012; Kamada et al., 2014; Martin et al., 2014; Peng et al., 2014;). Wang et al. (2012) have proposed a *de novo* assembly approach, processing separately SOLiD, 454, and Illumina datasets and forming supercontigs in a second step. The hybrid assembly with three platforms was significantly more efficient compared to each individual assembly (Wang et al., 2012). The integration of Illumina and 454 sequencer data is relatively easy because all reads are nucleotide sequences (base-space), unlike the SOLiD platform that produces dibase-encoded reads known as colour-spaces. SOLiD sequencing chemistry is based on the use of ligase and the extension of dibases at each cycle, and each base is sequenced twice (Metzker, 2010). This platform is consequently more efficient at discriminating sequencing errors and true polymorphisms (Metzker, 2010). The major limitation is the complexity of processing all reads in a colour-space bioinformatic pipeline; the translation of colour-space to base-space of read datasets produce a high proportion of aberrant read sequences (Ondov et al. 2008; Liu et al. 2012, Marco and Griffiths-Jones, 2012). The integration of SOLiD datasets with those from other platforms will be in a colour-space system, as suggested by Salmela (2010). This study describes a system of correcting sequencing errors, combining reads from SOLiD and other platforms in colour-space, and increasing the reliability of bioinformatic analysis, especially for reads with low sequencing coverage (Salmela, 2010). We describe a new approach for *de novo* assembly, integrating SOLiD and Illumina read datasets in a colour-space system for improving transcriptomic prediction. This hybrid colour-space read-additive (HCRA) approach was tested with simulated SOLiD and Illumina RNA-seq runs, compared to classical *de novo* transcriptomic assembly approaches, and applied to black pepper (*Piper nigrum*) RNA-seq datasets.

Results and Discussion

Generation of simulated dataset and transcriptomic assemblies

To better assess the quality of the CA and HCRA approaches, we generated a simulated dataset from the model plant *Arabidopsis thaliana* composed from an Illumina run, with 12.5 million paired-end reads, and one SOLiD run, with 25.5 million single-end reads. A range of k-mers was tested for assembly with Velvet and Oases, ranging from 19 to 55 with different steps. Those that achieved the best results are described in Materials and Methods. Automatic word size was used for CLC.

Comparison of single and multiple k-mers

Table 1 shows the assembly results for the various processing methods. The SOLiD and Illumina runs were assembled separately with CLC software and Velvet/Oases tools. The numbers of transcripts obtained using CLC methods were only about 40 and 48% of those obtained by Velvet/Oases for the SOLiD and Illumina runs, respectively. The same results were observed for N50 and total size of the assembled transcriptome; the Velvet/Oases results were superior to those for CLC. The Velvet/Oases improved more with the SOLiD run, detecting 46% of CEGMA conserved proteins compared to 15% for Illumina. CEGMA is a computational tool based on comparison of a query sequences dataset with 458 highly conserved eukaryotic proteins and consequently reports the completeness of transcriptome prediction. This first result agreed with that by Surget-Groba and Montoya-Burgos, demonstrating that multiple k-mers are more efficient for *de novo* assembly (Table 1). This method also had a greater advantage for SOLiD runs.

Comparison of CA and HCRA methods

A further step was to compare the different approaches integrating read datasets from the Illumina and SOLiD platforms. The CA approach is the classical approach for which CLC and Velvet/Oases tools have been tested. The Oases-based pipeline also performed better, with 40% better N50, 46% more transcripts, and 61% higher total Mb. The reads mapped back to transcripts (RMBT) and percent of CEGMA proteins detected, however, were more similar, with only 10% of proteins and 2-3% of mapped reads superior for the Oases than the CLC tools (Table 1). The HCRA method is a new approach based on the integration of reads and not of contigs after a primary assembly. All read assembly was done in colour-space, and only the obtained transcripts were converted to base-space in the final step. The transcript number was lower than that obtained by the CA approach, but the N50 was much higher. The RMBT were also higher, with a larger observed difference for CEGMA proteins. The HCRA method allowed the reconstruction of up to 98% of sequences. The HCRA method, using the Velvet multiple k-mer method, integrated datasets better, especially integrating more reads and forming longer transcripts (Table 1). The CAP3 tool was used to optimize transcripts for combining the transcripts obtained by the HCRA and CA approaches. The result was a higher N50 and a large decrease in transcript number, thereby removing redundancy. The percent of detected CEGMA proteins was constant and equal to that in the HCRA method (Table 1). The number of *de novo* assembled reconstructed transcripts are presented for each processing method in Fig. 1. The number of reconstructed transcripts with homology to initial cDNA sequences was relatively similar between conditions, but the total number of assembled and full-length transcripts with homology was highly variable. The CA method had more transcripts, resulting in a data set of 225000 transcripts with 16% (35960 transcripts) with homology to initial cDNA sequences, but the proportion of full-length transcripts was low at about 33%. HCRA had a 3-fold higher percentage of full-length transcripts, and the total number of transcripts was lower, indicating a better transcriptomic prediction. Combining the transcript datasets of the CA and HCRA methods with CAP3 had the advantage of drastically decreasing the total transcript number but slightly decreasing the number of transcripts with homology, compared to the HCRA dataset (Fig.1).

Table 1. Assembly statistics for CLC and Velvet/Oases.

Assembly	Total (Mbp)	No. Transcripts	N50 (bp)	RMBT (%) SOLiD	RMBT (%) Illumina	CEGMA (%) Complete/Partial
CLC SOLiD	30	67585	448	42.56	25.03	31.45/64.52
CLC Illumina	34	56234	678	38.44	30.68	48.39/86.69
Oases SOLiD	113	166472	851	45.06	33.50	58.47/78.63
Oases Illumina	84	114797	966	36.52	30.02	56.85/80.65
CLC CA	63	121782	570	50.84	37.39	57.26/89.92
CLC HCRA	41	49884	1177	55.47	48.58	81.05/97.58
Oases CA	163	225811	931	51.27	40.29	63.31/86.69
Oases HCRA	209	172835	1617	59.42	58.30	98.39/99.60
CAP3 (HCRA + CA)	95	75757	1704	60.20	58.68	98.39/99.19

Abbreviations: bp, base pair; Mb, megabase pair; RMBT, reads mapped back to transcripts; CEGMA, percentage of complete and partial conserved genes identified using CEGMA (complete refers to those predicted proteins in the set of 248 with alignment lengths 70% of the protein lengths; if a protein is not complete but exceeds a minimum alignment score, then it is called a partial protein); CA, contig additive; HCRA, hybrid colour-space read-additive; CAP3, removing redundancy with CAP3 using assemblies CA and HCRA

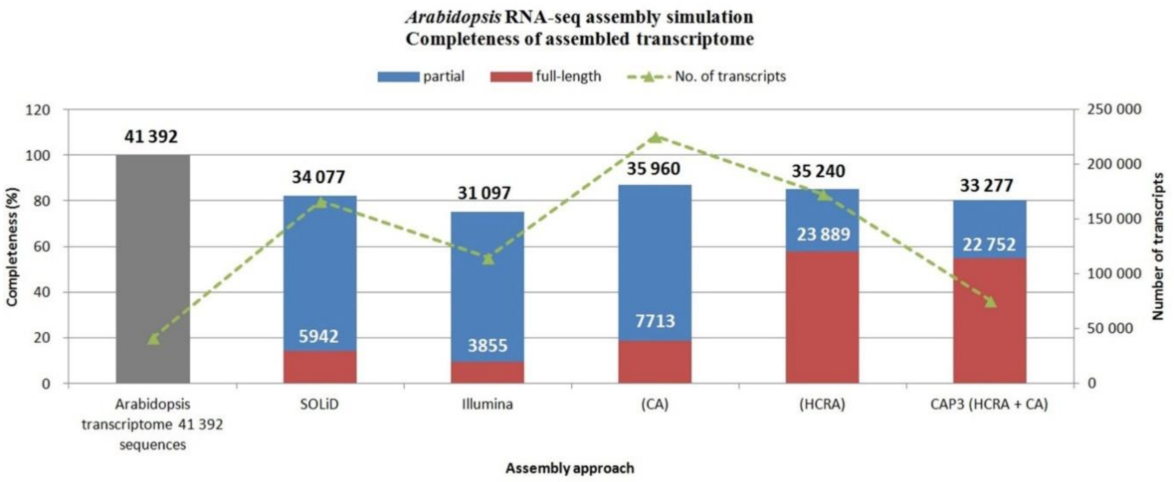

Fig 1. The percentage of unique transcripts partially and completely identified with Oases using A. thaliana reference cDNA. The blue bars represent partially recovered transcripts, red bars represent completely recovered transcripts, and the green line represents the number of transcripts for each approach.

Performance with a real dataset

A summary of our approach using *P. nigrum* data relative to simulated *A. thaliana* data is presented in Table 2. Our approach produced 60645 unigenes, totalling 71 Mb, with an average length of 1172 bp and an N50 of 1653 bp. Joy et al. (2013) reported 128157 unigenes, totalling 28 Mb, with an average length of 449 bp. Comparing our metrics to those by Joy et al. (2013), our method had 2.5-fold more total number of base pairs, 2-fold lower total number of unigenes, and an average of 2.5-fold longer unigenes. Gordo et al. (2012) reported 10338 unigenes, totalling 1 Mb, with an N50 of 168 bp. Comparing our metrics to those of Gordo et al. (2012), our method had 6-fold more unigenes, 70-fold more total number of base pairs, and a 10-fold higher N50. These results show that we could generate a more robust assembly and efficiently integrate data from SOLiD and Illumina. The percentage of the genes identified with CEGMA was >90%, completely and partially, and the percentage of RMBTs from Illumina was 44.10%. The low percentage of SOLiD RMBT can be explained by the low quality of the SOLiD sequencing, as illustrated by Gordo et al. (2012), which may also have influenced the decrease of 5% of the genes completely identified with CEGMA relative to *A. thaliana*.

Materials and Methods

Simulation of RNA-seq runs usingA. thaliana data

Simulated RNA-seq runs were generated by ART (Huang et al., 2012), version ART-VanillaIceCream-03-11-2014, using *A. thaliana* cDNA sequences from the ENSEMBL database. ART generates simulated reads by emulating the sequencing process with built-in, technology-specific error models and quality profiles parameterized empirically in large data sets (Huang et al., 2012). The SOLiD run is a single end 50 bp long and sequencing coverage of about 20×, and the Illumina run is a paired-end read of 100 bp, sequencing coverage of about 20×, and a pair distance of 300 bp. The simulated sequencing error rate for both runs was the default.

Contig additive de novo assembly

Velvet version 1.2.10 followed by Oases version 0.8.08 were used with a multiple k-mer method described by Surget-Groba and Montoya-Burgos (2010), and CLC version 6.5.1 was used with a single k-mer method. For CLC, the following parameters were used for the SOLiD and Illumina

Table 2. Assembly statistics of the HCRA+CA assembly for *A. thaliana* and *P. nigrum* data.

	A. thaliana	*P. nigrum*
Total Mb	95	71
No. of transcripts	75757	60645
N50 (bp)	1704	1653
Average length (bp)	1266	1172
Longest transcript (bp)	16326	12272
CEGMA (%) Complete	98.39	92.74
CEGMA (%) Partial	99.19	97.18
RMBT (%) SOLiD	60.20	25.65
RMBT (%) Illumina	58.68	44.10

Abbreviations: bp, base pair; Mb, megabase pair; RMBT, reads mapped back to transcripts; CEGMA, percentage of complete and partially conserved genes identified using CEGMA.

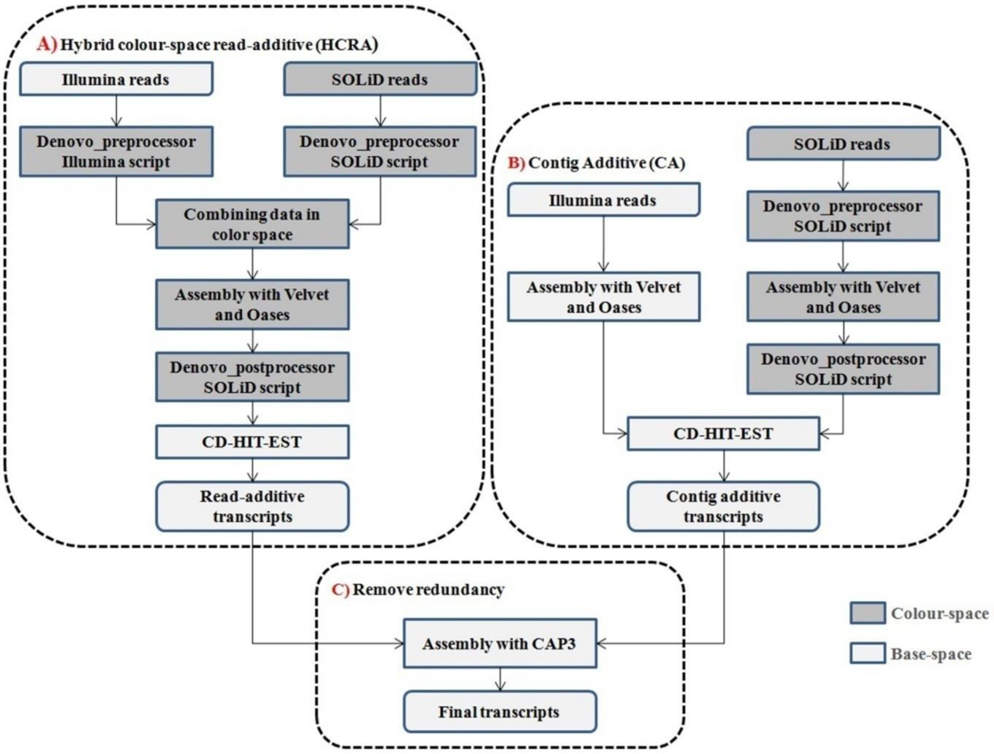
A) Hybrid colour-space read-additive (HCRA)
Illumina reads
SOLiD reads
Denovo_preprocessor Illumina script
Denovo_preprocessor SOLiD script
Combining data in color space
Assembly with Velvet and Oases
Denovo_postprocessor SOLiD script
CD-HIT-EST
Read-additive transcripts
B) Contig Additive (CA)
SOLiD reads
Illumina reads
Denovo_preprocessor SOLiD script
Assembly with Velvet and Oases
Assembly with Velvet and Oases
Denovo_postprocessor SOLiD script
CD-HIT-EST
Contig additive transcripts
C) Remove redundancy
Assembly with CAP3
Final transcripts
Colour-space
Base-space

Fig 2. The flowchart for *de novo* assembly using hybrid colour-space read-additive, contig additive, and combined approaches. (A) The hybrid colour-space read-additive approach. In this approach, Illumina reads are converted to SOLiD format with the denovo_preprocessor_illumina.pl script and grouped to SOLiD reads, the set of reads is submitted for assembly with Velvet/Oases, and the output is converted to base-space by the denovo_postprocessor_solid.pl script. (B) The contig additive approach. Illumina and SOLiD reads are assembled separately and the outputs are combined with CD-HIT-EST. (C) The two assemblies (CA and HCRA) are combined into a single set with CAP3 responsible for removing redundancy and extending the size of the transcripts.

runs: automatic bubble size, yes; minimum contig length, 200 bp; perform scaffolding, no; and k (automatic word size), 22.
Before assembly with Velvet and Oases, the SOLiD reads were converted to a double-encoded format using the "denovo_preprocessor_solid.pl" script from Life Technologies. After conversion, the SOLiD and Illumina reads were run separately for Velvet and Oases with: k-mer (k), 19- 25; step, 2; minimum size contig, 200 bp; perform scaffolding, no; and coverage, auto. After assembly, the SOLiD contigs were converted to base-space format using the "denovo_postprocessor_solid.pl" script from Life Technologies. In the second step, CD-HIT-EST (Li and Godzik, 2006) version 4.6 was used to combine assemblies with 100% identity. The contigs from assemblies CLC SOLiD and CLC Illumina were first combined, termed CLC CA, and then the contigs from Oases SOLiD and Oases Illumina, termed Oases CA.

Hybrid colour-space read-additive de novo assembly

The reads from SOLiD and Illumina required conversion before assembly with Velvet and Oases. The Illumina reads were first converted to colour-space and paired-end SOLiD format and then to double-encoded format using the "denovo_preprocessor_illumina.pl"script developed by us. The SOLiD reads were converted to a double-encoded format with the "denovo_preprocessor_solid.pl" script from Applied Biosystems. After the conversion, the SOLiD and Illumina reads were combined and subjected to Velvet and Oases with the parameters: k-mer, 19-25; step size, 2; minimum size contig, 200 bp; perform scaffolding, no; and coverage, auto. After assembly, the contigs were converted to nucleotides (conversion from double-encoded to base-space format) using the "denovo_postprocessor_solid.pl" script.

This approach is termed Oases HCRA. The SOLiD and Illumina reads were submitted to CLC without conversion and with the assembly parameters: automatic bubble size, yes; minimum contig length, 200; perform scaffolding, no; and k, 23.This approach is termed CLC HCRA. After assembly, the CLC HCRA and Oases contig datasets were submitted to the CD-HIT-EST tool with 100% identity to remove redundancy.

Merging assemblies for improved reliability

The Oases HCRA and Oases CA assemblies were combined with CD-HIT-EST to reduce the number of redundant transcripts, with 95% identity and using the default parameters of CAP3 (Huang and Madan, 1999) version 8.6.13. Fig. 2 shows a flowchart of assemblies HCRA, CA, and merging HCRA and CA with CAP3.

Quality assessment of transcriptomic prediction

Bowtie (Langmead et al., 2009) version 1.0.0 was used to evaluate the percentage of RMBT with default parameters: n (maximum mismatch number on seed), 2; l (size of seed), 28; and best parameter, yes.

The set of assemblies were compared to *A. thaliana* cDNA, available on the ENSEMBL database. BLASTN (Altschul et al., 1997) version 2.29 was used for comparison with the parameters: e-value, 1e-10 and max_target_seqs, 1. The script “analyze_blastPlus_topHit_ coverage.pl” (available at http://trinityrnaseq.sourceforge.net/analysis/full_length_transcript
_analysis.html) from the Trinity package analysed the full-length transcripts (transcripts covered by more than 90% of their transcript lengths.). CEGMA (Parra et al., 2007) was used to evaluate the completeness of the assembled transcriptomes.

Applying a pipeline for areal data set of P. nigrum sequences

We downloaded three runs of *P. nigrum* sequences from the NCBI SRA database. Two were single-end read runs 50 bp in length generated by the SOLiD 3.0 platform, one totalling 3.6 Gb (accession number SRX104901) (Gordo et al., 2012) and the other totalling 2.7 Gb (accession number SRX192196). The third was a paired-end Illumina HiSeq 2000 run (accession number SRX119532) with a read size of 100 bp and totalling 5 Gb (Joy et al., 2013).

After downloading the data, we applied the pipeline described in Fig. 2 to the real dataset, merging HCRA and CA and using the same parameters as for the simulated dataset. CD-HIT-EST was used to remove redundancy with 90% identity. The CEGMA and RMBT parameters were used to evaluate the completeness and percentage of incorporated transcripts in our approach, respectively.

Conclusions

SOLiD technology will likely soon be discontinued, but NCBI SRA contains more than 4500 SOLiD RNA-seq runs, and many more have certainly not yet been deposited and processed due to the complexity of colour-space read processing. The previously described studies (Salmela, 2010; Wang et al., 2012), however, have demonstrated that results are globally improved when SOLiD is combined with other platforms. Plant genomes are generally highly repetitive and often polyploid. Many plant species have no reference data, which complicates assembly. The *de novo* hybrid assembly of data from non-model plants, however, represents a viable solution for generating an initial set of high quality transcripts, as demonstrated by Agarwal et al. (2012) and Garg et al. (2011). The present study presents a hybrid colour-space read-additive (HCRA) approach to *de novo* assembly for non-model organisms that is able to reconstruct a data set of high quality, with data from the SOLiD and Illumina platforms using a multiple k-mer method. We applied HCRA and CA approach to a dataset of *P. nigrum* sequences available on NCBI SRA, generating about 60000 unigenes representing 71 Mbp of the transcriptome. These unigenes provide a valuable upgrade of the *P. nigrum* transcriptomic data and serve as a material basis for future genomic research of black pepper.

Acknowledgements

This work was supported by the Rede Paraense de Genômica e Proteômica (Governo do Estado do Pará/SEDECT/FAPESPA), PROPESP/UFPA, FADESP, FINEP and CNPQ.

References

Agarwal G, Jhanwar S, Priya P, Singh VK, Saxena MS, Parida SK, Garg R, Tyagi AK, Jain M (2012) Comparative analysis of kabuli chickpea transcriptome with desi and wild chickpea provides a rich resource for development of functional markers. PLoS One. 7(12):e52443.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389-402.

Garg R, Jain M (2013) RNA-Seq for transcriptome analysis in non-model plants. Methods Mol Biol. 1069:43-58.

Garg R, Patel RK, Jhanwar S, Priya P, Bhattacharjee A, Yadav G, Bhatia S, Chattopadhyay D, Tyagi AK, Jain M (2011) Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. Plant Physiol. 156(4):1661-78.

Gordo SM, Pinheiro DG, Moreira EC, Rodrigues SM, Poltronieri MC, de Lemos OF, da Silva IT, Ramos RT, Silva A, Schneider H, Silva WA, Sampaio I, Darnet S (2012) High-throughput sequencing of black pepper root transcriptome. BMC Plant Biol. 12:168.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7):644-52.

Huang W, Li L, Myers JR, Marth GT (2012) ART: a next-generation sequencing read simulator. Bioinformatics 28(4):593-4.

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res. 9(9):868-77.

Joy N, Asha S, Mallika V, Soniya EV (2013) De novo transcriptome sequencing reveals a considerable bias in the incidence of simple sequence repeats towards the downstream of 'Pre-miRNAs' of black pepper. PLoS One. 8(3):e56694.

Kamada M, Hase S, Sato K, Toyoda A, Fujiyama A, Sakakibara Y (2014) Whole genome complete resequencing of Bacillus subtilis natto by combining long reads with high-quality short reads. PLoS One. 9(10):e109999.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Adam MP (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30(7):693-700.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10(3):R25.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22(13):1658-9.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012:251364.

Marco A, Griffiths-Jones S (2012) Detection of microRNAs in color space. Bioinformatics. 28(3):318-23.

Martin JA, Johnson NV, Gross SM, Schnable J, Meng X, Wang M, Coleman-Derr D, Lindquist E, Wei CL, Kaeppler S, Chen F, Wang Z (2014) A near complete snapshot of the Zea mays seedling transcriptome revealed from ultra-deep sequencing. Sci Rep. 4:4519.

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12(10):671-82

McGettigan PA (2013) Transcriptomics in the RNA-seq era. Curr Opin Chem Biol. 17(1):4-11.

Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet. 11(1):31-46.

Ondov BD, Varadarajan A, Passalacqua KD, Bergman NH (2008) Efficient mapping of Applied Biosystems SOLiD sequence data to a reference genome for functional genomic applications. Bioinformatics. 24(23):2776-7.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 23(9):1061-7.

Peng Y, Lai Z, Lane T, Nageswara-Rao M, Okada M, Jasieniuk M, O'Geen H, Kim RW, Sammons RD, Rieseberg LH, Stewart CN, Jr. (2014) De novo genome assembly of the economically important weed horseweed using integrated data from multiple sequencing platforms. Plant Physiol. 166(3):1241-54.

Salmela L (2010) Correction of sequencing errors in a mixed set of reads. Bioinformatics. 26(10):1284-90.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 28(8):1086-92.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res. 19(6):1117-23.

Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res. 20(10):1432-40.

Wang Y, Yu Y, Pan B, Hao P, Li Y, Shao Z, Xu X, Li X (2012) Optimizing hybrid assembly of next-generation sequence data from Enterococcus faecium: a microbe with highly divergent genome. BMC Syst Biol. 6 Suppl 3:S21

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10(1):57-63.

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, Zhou X, Lam TW, Li Y, Xu X, Wong GK, Wang J (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 30(12):1660-6.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5):821-9.