scientific reports

Check for updates

OPEN

Raw transcriptomics data to gene specific SSRs: a validated free bioinformatics workflow for biologists

D. N. U. Naranpanawa[1,2], C. H. W. M. R. B. Chandrasekara[1], P. C. G. Bandaranayake[1] & A. U. Bandaranayake[3]✉

Recent advances in next-generation sequencing technologies have paved the path for a considerable amount of sequencing data at a relatively low cost. This has revolutionized the genomics and transcriptomics studies. However, different challenges are now created in handling such data with available bioinformatics platforms both in assembly and downstream analysis performed in order to infer correct biological meaning. Though there are a handful of commercial software and tools for some of the procedures, cost of such tools has made them prohibitive for most research laboratories. While individual open-source or free software tools are available for most of the bioinformatics applications, those components usually operate standalone and are not combined for a user-friendly workflow. Therefore, beginners in bioinformatics might find analysis procedures starting from raw sequence data too complicated and time-consuming with the associated learning-curve. Here, we outline a procedure for de novo transcriptome assembly and Simple Sequence Repeats (SSR) primer design solely based on tools that are available online for free use. For validation of the developed workflow, we used Illumina HiSeq reads of different tissue samples of *Santalum album* (sandalwood), generated from a previous transcriptomics project. A portion of the designed primers were tested in the lab with relevant samples and all of them successfully amplified the targeted regions. The presented bioinformatics workflow can accurately assemble quality transcriptomes and develop gene specific SSRs. Beginner biologists and researchers in bioinformatics can easily utilize this workflow for research purposes.

Abbreviations

CTAB	Cetyl trimethyl ammonium bromide
DBG	De Bruijn graph
GB	Giga bytes
Gb	Giga bases
NGS	Next generation sequencing
PCR	Polymerase chain reaction
SSR	Simple sequence repeat

Background

During the past decade, DNA and RNA sequencing technologies have made tremendous progress, in terms of throughput, speed and reduction of sequencing cost[1]. Similarly, access to genomes and transcriptomes have greatly benefited animal and plant biology research[2–7]. Sequencing technologies have evolved faster than one can expect. Second or Next Generation (NGS), and third generation sequencing technologies which are currently in use[8–10] (Table 1) boast a vast improvement from the first-generation sequencing technologies, especially with regard to throughput/cost ratio and speed. Traditional Sanger sequencing, which was widely used for almost

[1]Agricultural Biotechnology Centre, Faculty of Agriculture, University of Peradeniya, Peradeniya 20400, Sri Lanka. [2]Postgraduate Institute of Science, University of Peradeniya, Peradeniya 20400, Sri Lanka. [3]Department of Computer Engineering, Faculty of Engineering, University of Peradeniya, Peradeniya 20400, Sri Lanka. ✉email: asithab@eng.pdn.ac.lk

Sequencing instrument	Throughput
Illumina	
NextSeq	120 Gb[25]
HiSeq X	900 Gb[14]
Ion torrent	15 Gb[16]
SOLiD	155 Gb[17]
PacBio	50 Gb[19]
PromethION	15 Tb[22]

Table 1. Maximum throughputs recorded for sequencing platforms.

three decades since its publication in 1977, could only achieve limited or very low throughput[11]. The human genome project employed Sanger sequencing methods, and required over 10 years and nearly US$3 billion for completion[12]. In contrast, the Illumina HiSeq system can now sequence over 45 human genomes for US$1000 each in a single day[13]. Illumina is one of the NGS technologies where its HiSeq X instrument can achieve up to 900 Gb of total throughput at cost per Gb of a merely US$7[14]. While Illumina is considered the most popular NGS technology for short-read sequencing[14, 15], the second generation also includes other technologies such as Ion torrent and SOLiD by Life Technologies, where the throughput can reach up to 15 Gb[16] and 155 Gb, respectively[17]. Third generation sequencing methods such as SMRT sequencing by Pacific Biosciences (PacBio)[18] has the potential to produce around 50 Gb of data per SMRT cell with reads as long as 190 kb[19]. Nanopore sequencing by Oxford Nanopore can generate very long read-lengths[20,21], and their PromethION platform promises a yield up to 15 Tb of data in 2 days[22]. Further, it would only cost US$900 for a Nanopore platform to decipher around a billion DNA bases[23,24].

Due to these drastic advances, the amount of raw sequence reads produced by the sequencers are huge, and the high coverage adds up a massive amount of overlapping fragments of DNA/RNA, especially in large genomes[26]. Because the volume of data to be handled is very high, assembling the short reads back to construct the complete genome or transcriptome becomes challenging, requiring high computational power and execution time. This leads to a significant bottleneck in computational biology and bioinformatics[27].

Assembly of raw sequence data follows either of two approaches: (1) reference based[28,29] and (2) de novo assembly[30–35]. Reference based assembly, also called comparative assembly, is the process of recreating the genome or transcriptome using prior knowledge. In this method, a previously assembled genome of a closely related organism is used as a template to map or align the sequenced reads in question. Every read is placed at its most likely position against the reference assembly. The resulting assembly could be similar to the reference but not completely identical as there could be regions that are significantly different[36]. Therefore, comparative assembly is mostly used in genome re-sequencing projects and is considered a computationally easy task[37].

Assembling sequence reads with no prior knowledge of the transcriptome or without a reference genome is called de novo assembly. While de novo assembly provides the opportunity to assemble any novel organism, the process presents many challenges[38] including segmental duplicates, sequence repeats, missing or fragmented genes, and the massive amount of raw reads to be handled. Applying de novo assembly methods for plant genomes gives rise to even more limitations due to the size and complexity of plant genomes compared to animal genomes[39]. Furthermore, de novo assembly is mathematically proven to be difficult, given that it belongs to a family of problems with NP-hard complexity for which no efficient solution is known yet[40]. Nonetheless, de novo assembly is widely used over comparative assembly since many complex organisms are yet to be sequenced and closely related reference genomes are not always available[41,42].

While there are many applications and uses of assembled transcriptome of an organism, the identification of molecular markers[43,44] plays an important role in breeding programs that amplifies plant characteristics such as resistance and yield[45–47]. Of the different types of molecular markers, microsatellites[48]—also known as simple sequence repeats (SSR)—have been utilized most extensively. Transcriptome based SSRs have now replaced genome based SSRs because it is more effective and less expensive, and a number of such microsatellite markers have been published[49–52].

Even though sequencing technologies have advanced rapidly in a short span of time, methods and software used for assembly and analyses of sequence data[37, 53] have not seen the same degree of improvement. While most of these tools are still being revised for better algorithmic approaches and efficiency[54–56], the knowledge gap in bioinformatics has not allowed the rate of improvement to increase. One of the main reasons for this limitation lies in the fact that these assembly and annotation software are mostly commercial and very expensive[57]. However, free and open-source software play a major role in bridging this gap, not only by allowing anyone to test and experiment with their data, but also by allowing them to make changes and suggest improvements for the said tools. Even so, there is a lack of identified complete workflows built using such free software. This is again a setback for novice biologists as the workflow up to analysis of raw sequence data might contain several different procedures, and each individual tool might take up a considerable amount of time to decipher its workings.

Currently there are many free pipelines and frameworks available for transcriptome assembly using RNA-seq data. Galaxy[58] is a popular web-based online platform to build scientific workflows using preconfigured tools within. Considering it is a shared resource on public servers, the disk space quota and the number of concurrent analysis jobs an individual can have at a time are limited[59]. To address privacy and space issues Galaxy offers commercial clouds, which might prove to be expensive[60]. Even though Academic clouds are also

provided, configuration of the platform is complex. A complementary web-based tool to Galaxy is Taverna[61], which allows the integrating of third-party web-service tools to a Galaxy workflow. In addition to being fairly network intensive, Taverna also has limitations on the size of data it can handle. It is inherently targeted at a more expert audience and might require specific knowledge such as handling servlet containers and Java servlets. Hence, setting up, configuration and usage of both Galaxy and Taverna is not easy for a beginner with little to no computer science background. Open-source, cloud-based workflows are also available for RNA-seq analysis such as Arvados[62], Agave[63] and RAP[64]. However, one of the main disadvantages of these is that they impose a lot of constraints on the user as to the number of analyses that can be performed, and the size of raw data that can be uploaded to the cloud. In addition, they have retention policies where the data uploaded by the user will only be maintained at the cloud server for a limited amount of time after analyses have been performed. Cloud-infrastructure also demands high-speed internet connections for file transfer. Therefore, if such infrastructure is not available for the user, they might be discouraged from using such cloud computing environments. Bpipe[65] is a command line programming language that has been designed for defining and executing bioinformatics pipelines including transcriptome assembly. However, the user has to learn a completely new set of syntax and semantics specific for the Bpipe platform in order to use it efficiently. Nextflow[66] is another command line application that can be used for various RNA-seq analyses. However, the transcriptome assembly process of Nextflow is reference-based and cannot be used for de novo assembly. TransFlow[67] is a framework providing five independent modules that can be combined to build different transcriptome assembly workflows. Conversely, it requires the user to configure and specify parameters such as *kmer* lengths for assemblers for optimizations, which might be difficult for a beginner to decide at first.

In essence, most of the currently available frameworks for de novo transcriptome assembly are too complicated for a beginner biologist to start analyzing RNA-seq data immediately. Further, the outcome of most of these pipelines is only the transcriptome assembly, and does not extend into any further analysis. Conversely, there are a lot of tools available for post-processing and analyzing the assemblies with applications such as SSR marker designing and annotation[68,69]. However, majority of these do not provide any pipeline for de novo assembly from raw reads and expect the user to make available the assembled genome or transcriptome.

Here we present a complete workflow of free software that can be executed in a Linux environment. All the tools that we have used were developed by various other research groups and scientists, and are freely available online to download and execute (Supplementary File 01). Some of the software are even open-source, which allows users to suggest improvements and fix bugs, inherently improving the performance of the tools in return. Our workflow spans from acquiring sequenced reads, through quality control and assembly of data, up to assembly quality assessment and SSR primer design. For the current study, transcriptomic data was downloaded from a published study on *Santalum album* (sandalwood)[70]. We generated SSRs targeting few important oil biosynthetic genes with the objective of identifying makers for future breeding efforts. A batch of designed primers were validated with laboratory experiments.

Methods

Data acquisition. The majority of the sequencing data generated through sequencing projects are deposited at the Sequence Read Archive (SRA)[71] maintained by the United States National Institutes of Health National Center for Biotechnology Information (NIH/NCBI). The NCBI hosts a multitude of resources for bioinformatics and provides access to over 35 sequence databases[72] including GenBank and PubMed. GenBank also coordinates with repositories maintained by European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ). Between these primary databases, SRA contains more than 10,000 terabases of raw sequence data[73] as of 2018. NCBI also provides the SRA Toolkit software[74] freely. This can also be used to perform various sequence read file manipulation operations including downloading raw reads from NCBI and converting data file formats to suit various processing requirements.

For the current work, we used previously published transcriptomic data (BioProject PRJNA297453) of *S. album* (sandalwood) generated from four oil-producing sandalwood trees for a study of exploring the biosynthetic enzymes of key components of sandalwood fragrance[70], and the group isolated RNA from the tissues of three development stages of the trees; Sapwood (SW), Transition Zone (TZ), and Heartwood (HW). The *S. album* data were paired-end Illumina reads with a total of 117.98 Giga Bases in size.

The dataset (Supplementary file 02) was directly accessed through the FTP directory hosted by NCBI[75], and *prefetch* and *fastq-dump* commands of the SRA Toolkit were used to download each read file (Supplementary file 03). If the required dataset is small, Linux *wget* command can be used to download them directly. In contrast, the Toolkit can be used to download large datasets that span several libraries and experiments. After downloading, the SRA Toolkit was used to split the paired-end *sra* files into its respective forward and reverse read files in *fastq* format.

Quality control of data. After acquiring data, it is essential to identify basic statistics and the quality of the reads before proceeding into assembly and downstream analyses since low quality reads will result in low quality assemblies. We used the Linux command-line version of the FastQC package[76] to identify and evaluate the read quality by processing the *fastq* files (Supplementary file 04). FastQC is an open-source tool compatible with all main sequencing platforms. It allows the observation of read quality across all raw reads of a sample including diagnostics such as GC content distribution, average base quality per score, and adapter content. Reports generated by FastQC are supported by visualizing plots and accompanied by warnings about uncertain results.

After evaluating the quality based on the reports provided by FastQC, the reads that did not meet the defined standards were filtered using FASTQ quality filter in FASTX-toolkit[77] (Supplementary file 04). FASTX can remove reads or nucleotides that are below a certain threshold specified by the user based on the insights gained from

FastQC. If FastQC reports indicate high adapter content, scripts included in Trimmomatic[78] can be used to trim such adapter sequences introduced by NGS instruments when sequencing (Supplementary file 05).

In some specialized cases such as assembling RNA-seq data, additional steps can be taken to clean up the reads. RNA-seq data could include ribosomal RNA (rRNA), which could cause errors in downstream analyses if not removed at early stages. To clean such rRNA that might be present in the dataset, we used SortMeRNA[79] (Supplementary file 06). After performing read filtering, FastQC reports were generated again to assess if previously reported warnings have been resolved and if the read quality has improved.

De novo *transcriptome assembly*. We chose Trinity de novo assembler[80] for the assembly of the *S. album* RNA-seq dataset. Trinity is considered to provide high quality assemblies[81] and consists of three software modules: Inchworm, Chrysalis and Butterfly. These stages work in an automated flow resulting in a complete transcriptome assembly.

For paired-end reads, Trinity takes two fastq files as its input; a forward end read file and a reverse end read file. Since multiple RNA-seq files from different libraries were available in our dataset, all forward end reads and all reverse end reads were concatenated into two separate fastq files by using simple Linux commands (Supplementary file 08). Attention was given in setting parameters such as maximum memory to be used by Trinity (–max_memory), number of CPUs to use (–CPU), and normalizing reads (–no_normalize_reads). Memory and CPU usage depends on the hardware platform that Trinity has been installed on. Normalizing is usually required for large datasets with deep sequencing as redundant data could be present. We set default values for all transcriptome assembly parameters and set the dataset to be normalized with in silico normalization[82] provided by the Trinity pipeline (Supplementary file 07, Supplementary file 08). The assembly was run on a server with 32 cores and 128 GB of Random Access Memory.

Trinity assembles the raw reads into 'contigs'. A contig is the smallest assembly component and it represents sets of overlapping DNA that can be summed to form a contiguous region of DNA. For further analysis, these contigs are required to be clustered into larger sequences – called 'unigenes' within the Trinity environment. This process is internally carried out by Trinity without user intervention. If further processing is required, Trinity provides scripts to avoid redundant transcripts and filter the 'longest isoform genes' as well.

Assessing and validating assembly quality. As de novo assembly is performed without no prior information available, quality assessment is a critical step. Using a low-quality assembly would not only misrepresent the results of any proceeding experiments, but it would also lower the reliability, credibility and repeatability of any related analyses.

Initial quality assessment was done with a Perl script provided in Trinity to retrieve basic statistics about the assembly. This script was run on the *fasta* file containing the transcriptome assembly, which generated the values for total length of assembly, number of contigs assembled, GC content, and N50 statistics of contigs and unigenes.

Further quality assessments were done with several other software (Supplementary file 09). The Bowtie2 program[83] was used to align the input sequence reads back to the transcriptome assembly obtained with Trinity. From the list of quality values provided by Bowtie2 after alignment, the percentage of raw reads mapped back to the transcriptome was observed primarily to assess the assembly quality. In addition, BUSCO assessment tool[84] was used to perform an evolutionary measure of genome completeness by searching the assembly against a reference database. For the current study, 'eukaryota_odb9′ reference database was downloaded from the BUSCO website which was the database closest to the species of interest. The statistics from the validation include the percentage of complete orthologs (single copy and duplicates), fragmented orthologs, and missing orthologs as well. The value generated for the number of complete orthologs was given priority for assessing the assembly quality. Furthermore, TransRate program[85] was also used to produce a number of program-specific quality metrics to further validate the assembly. Attention was given to the percentage of good mappings and assembly score to assess completeness.

Designing SSR primers for identified genes. For this study, we selected eight predicted oil biosynthetic genes of *S. album*[86] and two control genes (rbcL and TUB1), and the coding sequences of the preferred genes were downloaded from NCBI in *fasta* format. Then, using Linux command-line ncbi-blast + tool (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast + /), a blast database was created for the local transcriptome assembly generated by Trinity, which acts as a subject for following queries. Each gene sequence was then individually queried against the local database via the *blastn* command of ncbi-blast + and the significant alignments were directed into a separate text file. Sequence IDs of most significant alignments for each gene were subsequently filtered from the text file, and the list of IDs were used to extract the aligned sequences from the transcriptome assembly in *fasta* format.

Some of the selected genes showed duplicate alignments to the transcriptome. To ensure that we choose unique sequences for primer design, we used the Basic Local Alignment Search Tool (BLAST) web interface of the NCBI, which provides the facility to query against a published genome (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The preferred genome could be mentioned in the search bar under 'BLAST Genomes' by the organism name, scientific name, or tax id. Here, we used the *S. album* (taxid: 35,974) genome published under BioProject PRJNA411901. Sequences that were earlier extracted from the transcriptome assembly, were individually queried against the genome to determine the best sequence for primer design. Candidates with a unique alignment and lowest E-value were selected.

The BatchPrimer3 (UC Davis Server)[87] and OligoAnalyzer[88] tools were used for detecting SSR markers and designing primers. On BatchPrimer3, default parameters were used with 'SSR screening and primers' as the primer type, and the chosen candidate sequence was the input. The minimum number of SSR pattern repeats for

di, tri, tetra, penta, and hexa nucleotide SSR types were specified. From the resulting list, primers were evaluated considering GC content, primer melting temperature (Tm) and product length. The selected forward primer and reverse primer were separately analyzed by using the web tool OligoAnalyzer by observing the characteristics such as hairpins, Tm and GC percentage.

The forward and reverse primers were again queried separately against the selected *S. album* genome to assure unique and specific amplification, and such primers were selected.

Laboratory validation of designed SSRs. We conducted wet lab experiments to assess the accuracy of the de novo transcriptome assembly as well as the SSR primer design process. DNA was extracted from bark samples following hexadecyltrimethyl ammonium bromide (CTAB) method[89]. PCR was carried out in a 25 µL reaction volume containing lx PCR buffer, 1.5 mM MgCl2, 200 µM dNTP (Promega, Cat No: U1515), 0.2 µM of each primer (Integrated DNA technologies, Singapore), 100 ng of DNA, 0.8 µM spermidine and 1 Unit Go Taq Flexi DNA polymerase (Promega, Cat No: M8295). The PCR cycle consisted of 94 °C of initial denaturation for 3 min, followed by 35 cycles of 94 °C for 1 min, annealing temperature ranging from 55 °C-58 °C for 30 s (depending on primer) and 72 °C for 30 s and final extension at 72 °C for 5 min. Amplified products were separated by electrophoresis (5 V cm-1) on 2% agarose gels and Safegreen (abm G108-G) stained gels were visualized and photographed using ChemiDoc XRS$^+$ system with Image Lab Software (version 6.0.1.34), where excitation source was UV trans Illumination and emission filter was a standard filter. Further the products were separated using 8% polyacrylamide gel electrophoresis for higher resolution. Sizes of the PCR products were estimated with a 100-bp DNA molecular weight marker (promega G2101). In addition to that, a similar PCR experiment was conducted with three *S. album* accessions to check the polymorphism of the SSR marker set and applicability in future breeding work.

A summarized view of the complete assembly workflow is presented in Fig. 1.

Results

Data acquisition and quality control of raw reads. NCBI data is in SRA format by default when downloading. If the raw data is paired-end, the *sra* files need to be converted and split into its respective forward and reverse read files in *fastq* format for use of processing tools. If the data is single-end the conversion can be direct without splitting.

On the other hand, the above step is not applicable if the researcher has generated own sequencing data. Depending on facilities produced by the service provider, sequenced data may directly be downloaded in *fastq* format from a storage hosting service such as a cloud, or from the sequencing instrument.

The RNA-seq dataset of *S. album* downloaded from NCBI had a complete volume of 117.98 Gbases and 53.42 GB for 21 SRA accessions. Number of raw reads in forward and reverse read files of each SRA accession are given in Fig. 2. When the raw reads were assessed with FastQC with default parameters, generated reports indicated a certain level of low-quality reads based on the 'Per base Sequence Quality', 'Overrepresented sequences', and 'Adapter Sequences' metrics. The number of raw reads after filtering low quality reads with FASTX-toolkit are indicated in Fig. 3. Filtered and trimmed read files indicated only good quality reads were remaining upon rechecking with FastQC.

De novo *transcriptome assembly*. Using Trinity, a total of 628,438,851 bases were assembled into 771,200 contigs. Default values for all parameters were provided for the execution of the assembly command. The generated contigs were further clustered within Trinity assembly process into 604,666 unigenes with a mean length of 561 bp and N50 value of 659 (Table 2). Total assembly execution time for Trinity was 167,335 s with Chrysalis stage taking a major portion of that time (160,011 s).

Quality assessment of the transcriptome assembly. While basic quality metrics do provide an idea about the contiguity of the assembly, the applicability of them might change according to the type of assembly and amount of data.

The basic statistics generated by Trinity indicated that a total of 628,438,851 bases were assembled as contigs with an N50 of 1405. From the contig bases, only 339,069,641 bases were further clustered into unigenes with an N50 of 659 (Table 2).

The length of the assembly and the number of contigs generated provide a straightforward insight into the success of the assembly. Lower the number of contigs, better the assembly – but all the contigs as a group should cover a majority of estimated assembly size for accuracy. Previous studies or flow-cytometry analysis may help to estimate the assembly size beforehand. For an example, previous studies suggest that a chloroplast genome assembly would normally be between 120 kb (kilobases) to 170 kb in length[90–92]. Hence, if basic statistics of a chloroplast genome exceeds or falls below that general range, it is best to be revised again. The same concept applies for transcriptome assemblies as well.

The GC content of a sequence represents the percentage of nucleotide bases that are either guanine or cytosine on a DNA or RNA molecule. DNA sequences also include adenine and thymine bases, while RNA has uracil instead of thymine. A gene-rich genome can be roughly identified with the GC content, as it is indicative of many protein-coding genes. Low percentages of GC might indicate a large amount of non-coding DNA in the genome[93]. Knowing the GC content of an assembly or a sequence is also important for downstream experiments such as polymerase chain reactions (PCR)[94] and primer design[95] as their annealing temperatures might have to be determined. A higher GC percentage indicates a higher melting temperature.

N50 is a weighted median statistic provided for assessing the contiguity of an assembly fragmented by contigs of different sizes. The value is defined as the minimum contig length required to cover half of the genome or

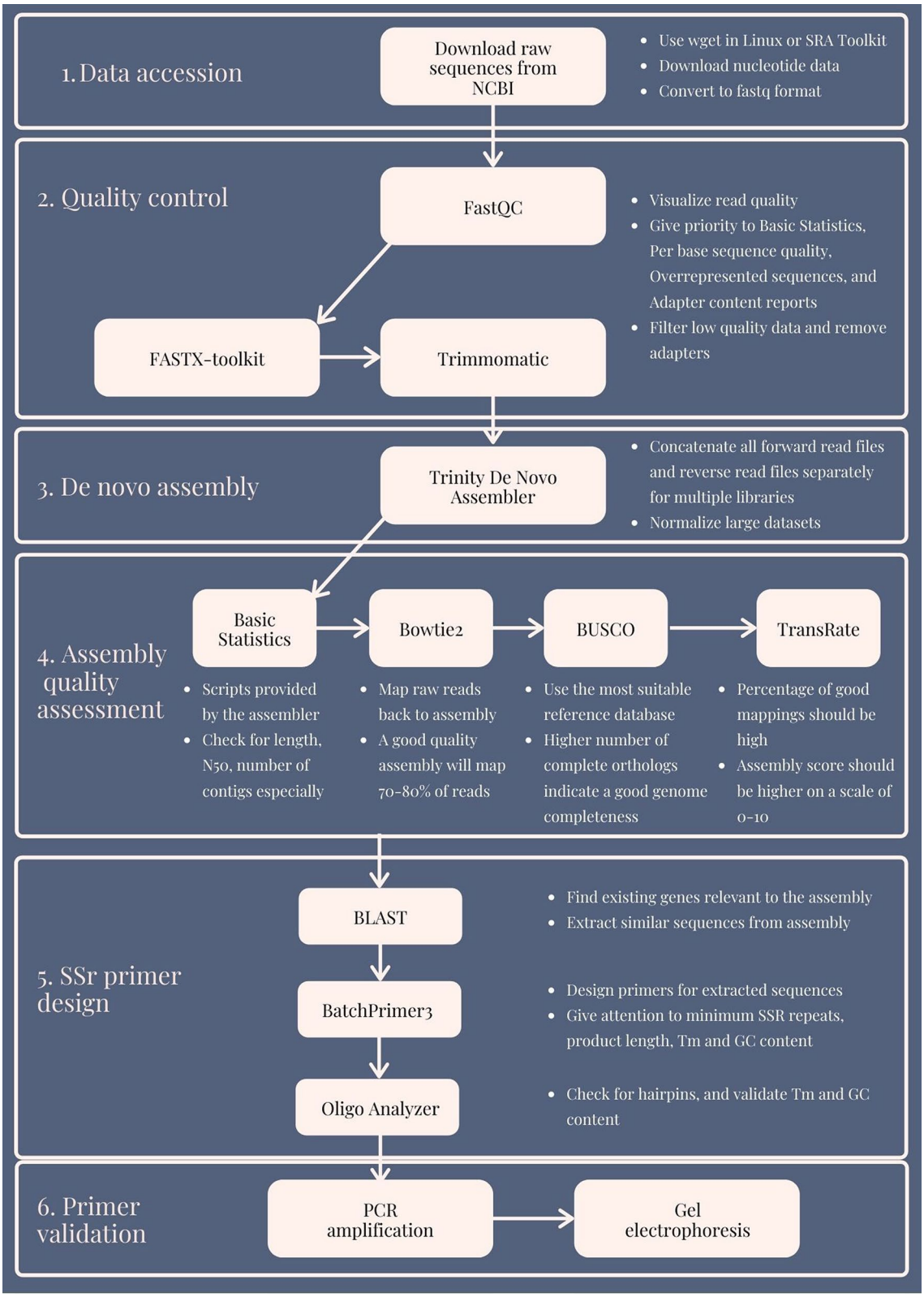

Figure 1. A summarized computational workflow for de novo transcriptome assembly, and quality assessment and assembly validation.

transcriptome[96]. That is, 50% of the sequence length is contained in contigs equal to or larger than the N50 value in length. Considering the total number of bases assembled (~628 Mb), the N50 of current analysis does not

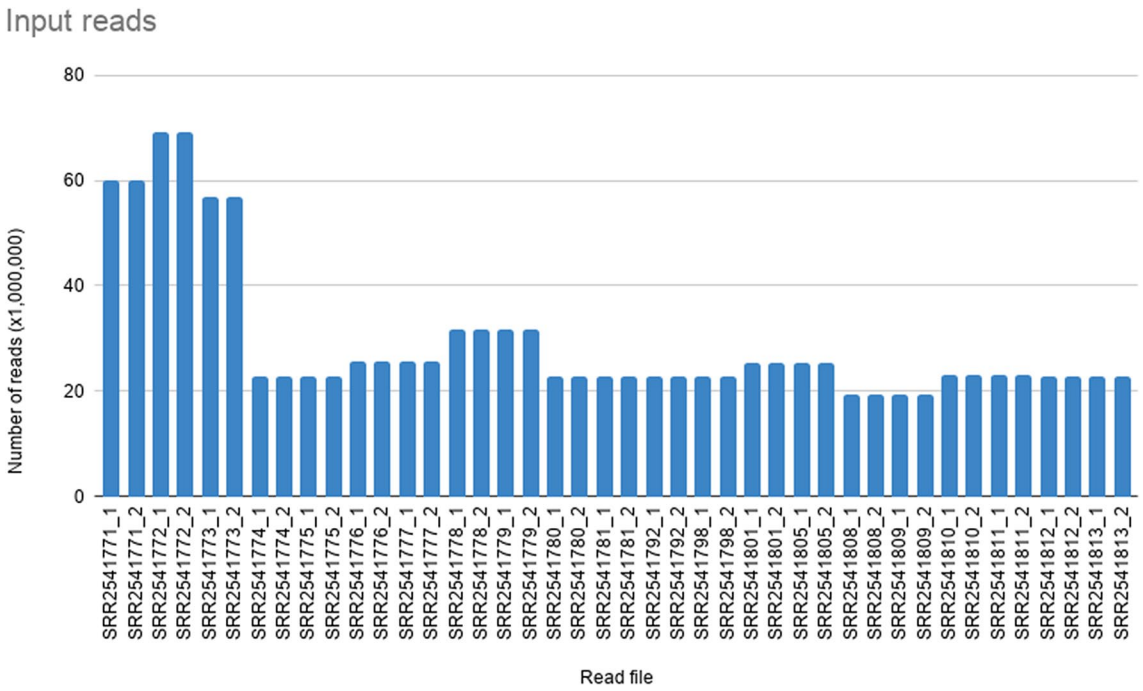

Figure 2. Variation in the number of raw reads in input RNA. Each accession file has a forward and reverse dataset indicated by suffixes _1 and _2 respectively.

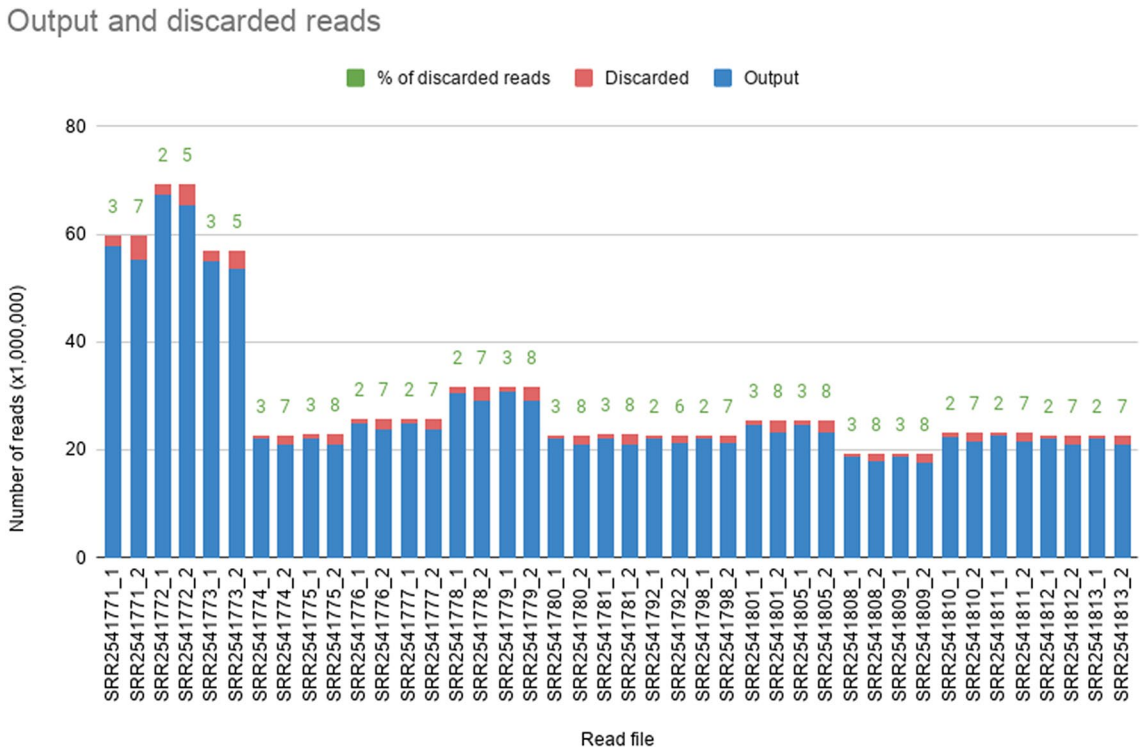

Figure 3. Variation in the number of raw reads in the output after filtering the original data with FASTX toolkit, and the number of filtered reads. Filtered reads as a percentage of input reads are also indicated. Each accession file has a forward and reverse dataset indicated by suffixes _1 and _2 respectively.

indicate a good quality assembly as it is considerably low. However, due to lowly expressed isoforms, N50 metric could behave in a biased manner[97]. In addition, even though N50 might convey a sense of scale and contiguity of the assembly, it does not correlate with the accuracy or the coverage of the assembly as demonstrated by recent large-scale assembly competitions[98–100]. In some cases, a large N50 value might be produced artificially due to large contigs that might have been misassembled[100]. Contigs might need to be corrected for erroneous

Description	Statistics
Sequence reads	
Raw reads	117.98 Gb
Assembly	
GC content (%)	39.13
Contigs	771,200
Number of bases	628,438,851
Mean length (bp)	814.88
N50	1405
Unigenes	604,666
Number of bases	339,069,641
Mean length (bp)	560.76
N50	659

Table 2. Summary statistics of sequence data and De novo assembly.

Read type	Count	%
Proper pairs	145,103,817	83.45
Improper pairs	17,730,957	10.2
Right only	5,553,993	3.19
Left only	5,495,262	3.16
Total aligned rnaseq fragments		173,884,029

Table 3. Summary of statistics of Bowtie validation.

BUSCO type	Count	%
Complete	278	91.7
Complete and single-copy	78	25.7
Complete and duplicated	200	66
Fragmented	21	6.9
Missing	4	1.4
Total BUSCO groups searched	303	

Table 4. Summary of statistics of BUSCO validation.

concatenations and N50 measured again for a more accurate indication of contiguity. All the same, N50 by itself cannot be guaranteed as a good measure for transcriptome assembly quality and does not necessarily indicate accurate contig orientation[98,101].

Bowtie2 execution on the transcriptome assembly indicated 83.45% of proper pairs out of the RNA-seq fragments aligned to the transcriptome (Table 3). Only 10.2% of the fragments were recorded as improper pairs.

Bowtie2 program[83] provides statistics as to how many cleaned reads actually represent the assembly. Assembly quality is high if at least 70–80% of sequence reads were mapped back to the assembly by Bowtie2[102]. The current assembly quality can be considered high since more than 80% of the reads indicated to be proper pairs. Given that the assembly was generated from many short reads as well as reads with average quality, this alignment is justifiable for the quality of the transcriptome.

BUSCO is a tool to assess completeness of genome assembly, gene set and transcriptome. It is based on the concept of single-copy orthologs that should be highly conserved among the closely related species. The BUSCO assessment reported that out of the 303 BUSCO groups searched using a global reference, 91.7% of the *S. album* transcriptome were complete orthologs, while only 1.4% were missing (Table 4).

The high percentage of complete single-copy orthologs generated by the BUSCO assessment tool[84] suggests a high quality and near-complete assembly since more than 90% of complete orthologs were present. The higher number of duplicated orthologs out of the complete orthologs indicates that multiple copies of full-length orthologs are found in the assembly. Since this is a transcriptome assembly, where multiple sequences are reconstructed at varying levels of abundance, this metric can be considered normal.

The TransRate assembly score was also considerably low at only 0.0105 and out of all RNA-seq fragments, only a 0.16% of fragments can be considered as good mappings (Table 5).

Read mapping metrics	Value
Fragments	181,844,966
Fragments mapped	31,616,682
Fragments mapped %	0.17
Good mappings	28,474,573
Good mappings %	0.16
Transrate assembly score	0.0105
Transrate optimal score	0.0618
Good contigs	318,291
Good contigs %	0.41

Table 5. Summary of statistics of TransRate validation.

	GenBank Accession	Definition	SSR Motif	Primer Sequence 5′ to 3′		Product size
1	KC842188.1	cytochrome P450 reductase (CPR2)	(GTGC)2	Forward	ATGCCCTCTGTTTAAGCTACT	151
				Reverse	GAACAGAGTCAATCAGATCGT	
2	KT160233.1	SaCPR722 cytochrome P450 reductase	(GGAAA)2	Forward	CAGCGAGGTTTTATAAATGG	146
				Reverse	CTCAGAAAGAATGTCATCCAC	
3	KT160234.1	*S. album* isolate SaCPR3442 cytochrome P450 reductase	(ACAG)2	Forward	CAGCGAGGTTTTATAAATGG	146
				Reverse	CTCAGAAAGAATGTCATCCAC	
4	KT160235.1	*S. album* isolate SaCPR7351 cytochrome P450 reductase	(TCGGC)2	Forward	AGTGGACTACGAGGATGAGTT	152
				Reverse	CATGAACATCACGAAACCTAC	
5	KT160236.1	*S. album* isolate SaCytB5-4631 cytochrome b5	(TTTCTT)2	Forward	CTCTCTCTCGATTCTGTTGTG	162
				Reverse	TATTTACTGGGCACACCTATG	
6	KT160237.1	*S. album* isolate SaCytB5-6548 cytochrome b5	(ATGG)2	Forward	ATAACGCTTCAGGAATAGGAC	137
				Reverse	CCCTCTGCTGTTAAAGATGAT	
7	KT160238.1	*S. album* isolate SaCytB5-3125 cytochrome b5	(TAGTA)2	Forward	TTCTCAATCCTTAGACCCACT	135
				Reverse	GACGGTTCAGATGCAAGTAT	
8	KT160239.1	*S. album* isolate SaCytB5-5956 cytochrome b5	(CACAA)2	Forward	AAGTTCCGTTCTCTCTGAATC	143
				Reverse	GAGGTTAGATTGTAAACCTTCC	
9	rBcL	ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit [*Arabidopsis thaliana* (thale cress)]	(AGTTCA)2	Forward	GTCCGATGGGATAGACTAAAA	159
				Reverse	GTTCAACCAACCCATTTTC	
10	TUB1	tubulin beta-1 chain [*Arabidopsis thaliana* (thale cress)]	(TCAA)2	Forward	GGTTGAGATCACCAACTGTAA	146
				Reverse	CCTTATGAAACATGCTTTGG	

Table 6. SSR motifs and primers designed for 10 gene sequences.

Therefore, in contrast to validations performed by both Bowtie2 and BUSCO, TransRate indicated poor assembly quality. The percentage of good mappings and assembly score indicates the level of assembly completeness[85]. The assembly score ranges from 0 to 10 according to the specifications of TransRate, and a higher assembly score would indicate a higher quality transcriptome assembly. Previous studies indicate that there is a threshold of 0.22 in the TransRate score which was achieved by only 50% of the 155 published de novo transcriptomes in NCBI[85]. The resulting score for current assembly is below this threshold. However, as in the case of the N50 metric, TransRate metrics could also be biased against a large number of lowly expressed transcripts, leading into a lower score[97]. More sequence data or information might be required to correct or improve these TransRate scores.

Validation of designed SSRs with PCR amplification. SSR primers that were designed for selected sandalwood genes and controls are presented in Table 6.

Using command-line ncbi-blast + tool on the Linux system seemed to be highly efficient than using the NCBI website for blastn queries. The NCBI web tool depends on the available network bandwidth and is congested frequently. In addition, the process would take too long, and saving results for later use is arduous. Instead, by using the Linux application for blast executions and simple Linux commands, files can easily be manipulated into providing succinct, exact outputs.

The *S. album* genome published in NCBI (BioProject PRJNA411901, Master accession NXEK01000000) was used to observe unique alignments of candidate sequences. When designing primers for these sequences on BatchPrimer3, minimum repeats for di, tri, tetra, penta, and hexa SSRs were set as 2, 2, 3, 3, and 3 respectively. From the results, eight pairs of forward and reverse primer sequences were tested unique against the *S. album* genome on NCBI (Table 6).

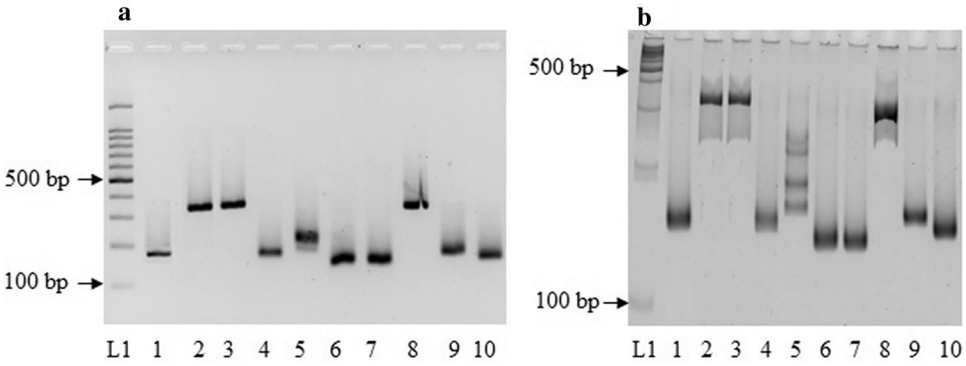

Figure 4. Polymerase chain reaction amplification of Simple Sequence Repeat markers and two housekeeping genes of *S. album*. (A): Agarose gel electrophoresis, (B): Polyacrylamide gel electrophoresis (PAGE). L1:100 bp molecular weight marker (promega G2101), 1:KC842188.1, 2:KT160233.1, 3:KT160234.1, 4:KT160235.1, 5:KT160236.1, 6: KT160237.1, 7:KT160238.1, 8:KT160239.1, 9:rBcL, 10 :TUB1. Full length gel image is presented in the Supplementary file 10.

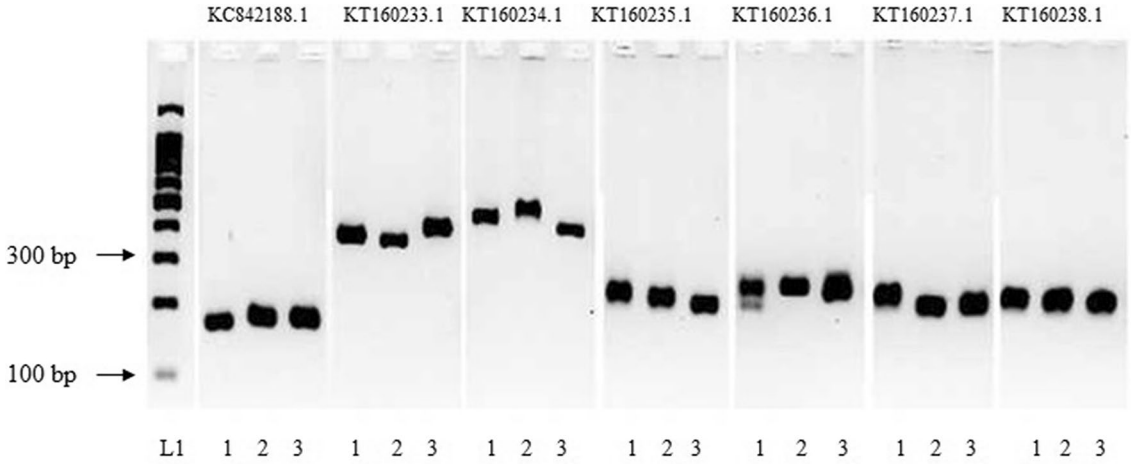

Figure 5. Agarose gel electrophoresis of Simple Sequence Repeats amplified products of three Sri Lankan *S. album* accessions. L:100 bp molecular weight marker (promega G2101), 1–3: *S. album* accessions. Full length gel image is presented in the Supplementary file 11.

All the eight SSR primer pairs amplified expected fragments from genomic DNA of selected *S. album* accession including control primers, rbcL and TUB1. While KC842188, KT160235, KT160237, and KT160239 amplified at 55 ℃, others amplified at 59 ℃. All the SSR markers except KT160236 appeared as single alleles in both the agarose and polyacrylamide gels, suggesting these novel markers possess a specific amplification in *S. album* (Fig. 4). However actual product sizes were greater than the expected sizes for KT160233, KT160234 and KT160239.

We further tested the polymorphism of selected primers with the objective of optimizing them for future breeding efforts. Of them, KT160233, KT160234, KT160235 and KT160237 resulted length polymorphism among three selected accessions (Fig. 5). This suggested high polymorphism among *S. album* accessions, as well as the ability to detect such variations with primers designed.

Discussion

We presented a complete workflow of free software from acquiring sequenced reads, through quality control and assembly of data, up to assembly quality assessment and SSR primer design. All the tools used are freely available online to download and execute while some of the software are open-source.

Quality control of raw reads is essential because sequencing instruments could introduce impurities during the sequencing process which would propagate into the final data output. In addition, sequencing platforms still suffer from various issues despite their rapid evolution[103]. They might not produce reads that are 100% aligned with the client requirements due to platform-specific biases[104], and might generate different types of errors in read data in varying amounts such as low quality bases and PCR errors[105,106]. For an example, some Illumina sequencing platforms have a tendency to produce reads that have lower base quality at the beginning and towards

the end of the read[107]. Since most tools in downstream analyses are not capable of detecting these errors during run-time, initial cleaning and filtering of the raw sequence reads is critical to obtain accurate results in analyses.

Several tools have been developed to identify and evaluate the read quality. Such tools can visualize base quality scores, nucleotide distributions, contamination and GC bias. Among them, FASTX-toolkit[77], which consists of several Linux command line tools, has been widely used. FASTX-toolkit is capable of checking base quality and nucleotide distribution of short-read FASTQ files. However, FASTX-toolkit only supports Illumina reads. Further, FastQC package[76] can also be used to generate insights about the raw reads. The SortMeRNA[79] tool can also clean rRNA from data generated from several platforms including Illumina, IonTorrent and PacBio.

However, if the read quality is extremely poor, re-sequencing would be the best option, since filtering alone would not correct such extreme errors. In addition, these errors would further escalate as the data proceeds along the workflow. Nevertheless, quality control will allow the identification of bad samples at early stages, reducing the amount of time spent on analyzing the data at later stages of the workflow. This would result in higher accuracy in the post-processing of data.

While a plethora of assemblers have been developed for assembly processes, it is important to choose an assembler that fits the requirements of the research. For an example, most genome assemblers are not optimized to carry out transcriptome assembly[108], and reference-based assemblers are not optimized for de novo assembly. Here we used Trinity de novo assembler[80], since our data were short-read paired end RNA-seq reads of a non-model organism. Trinity is considered to provide high quality assemblies[81].

Other popular software for de novo assembly of short-read RNA-seq data includes SOAPdenovo-Trans[108], Oases[109], and Trans-ABySS[110] which have been successfully applied in assembling the transcriptomes of various organisms[30,31,111,112]. All of these assemblers use de Bruijn graphs (DBG) to construct the transcriptome assembly. The DBG method is most popularly employed for short-read assembly, and uses a form of K-mer graphs to build the assembly from raw reads[113].

SOAPdenovo-Trans is derived from SOAPdenovo2 genome assembler[114], developed as a solution for the algorithmic challenge of assembling very short paired end RNA-seq reads into complete or full-length transcript sequences. Similarly, Oases was extended from Velvet assembler which was originally developed for genome assembly. Trans-AbySS addresses variations in local read densities. Trinity was developed by the Broad Institute in collaboration with the Hebrew University of Jerusalem and is a good general solution for de novo assembly as well as genome-guided transcriptome assembly. It applies three software modules sequentially on large volumes of RNA-seq data, and correct transcript reconstruction is ensured since it was specially programmed to remove ambiguous and erroneous reads. Further, Trinity has assembly options for both single-end and paired-end reads.

Trinity also provides the option of normalizing the reads before assembling. Normalizing is usually required for large datasets with deep sequencing, as redundant data could be present[115]. This would efficiently reduce the computing requirements for de novo assembly including time and memory requirements while retaining the sequencing coverage[116]. Nevertheless, not all datasets require normalization. If enough computing power is available, smaller datasets with lower coverage can be assembled without normalizing, as normalization would discard only the reads that are above a high coverage threshold[117]. Therefore, it is important to correctly identify the transcriptome complexity as well as read properties before normalizing data. We recommend performing assembly both with and without normalization, and evaluate the assembly quality to identify the best course of action. While stand-alone normalization tools such as the khmer software package [118] are available, Trinity conveniently provides in silico normalization[82] in the pipeline itself. Trinity also bundles Trimmomatic with it, enabling the user to trim and correct low quality reads and adapter sequences. Trinity normalizes, trims, and assembles consecutively with one command with necessary input parameters.

By a majority of the results, the *S. album* transcriptome assembly can be considered to be of good quality. However, the discrepancy between the metrics generated by various evaluation tools shows that further assessments are needed to validate the assembly. Such discrepancies are not uncommon[119] and reasons are discussed by several authors[56,99,100] . The original study from which we obtained data had followed a somewhat different methodology in preparing the sequence reads for the transcriptome assembly, but had also used Trinity for assembly[70]. The mean length of the transcripts of their assembly was 864 bp. The transcriptome we assembled had transcripts with a mean length of 814.88 bp, which indicates the assemblies were similar in quality. One of the options to improve the overall quality would be to re-sequence with different specifications.

Nevertheless, the expected quality depends on the downstream applications. For example, if the assembly is used for read mapping and differential expression analysis, the initial set of transcripts might be sufficient[120]. For genetic diversity studies and evolutionary assessments[121], high quality assemblies might be required. However, an assembly with extremely poor quality should not be used for any downstream analyses as it might not represent the genomic or transcriptomic information about the organism accurately[122]. Simple errors can accumulate into significant mis-assemblies[123]. These might result in inaccurate reconstructions of the genome/transcriptome, leading to false results and conclusions.

SSR primer design is one of the downstream applications of assembled transcriptomes. Here we considered several oil biosynthesis genes of *S. album* with the objective of developing gene specific markers for future breeding efforts. While all the primers tested resulted clear PCR products, the product sizes were greater than the expected sizes for KT160233, KT160234 and KT160239. The larger size may attribute to genetic differences in number of repeats in the SSR motif between different accessions. While the RNA-seq data was from an Australian *S. album* accession, validation was done with a Sri Lankan accession. Further, multiple alleles of KT160236 present in the selected accession did not appear in the bioinformatics analysis. Such kind of polymorphism is common in SSR motifs[98].

Interestingly, the genetic polymorphism observed is correlated with the related chemical constituents of the genotype (data not shown), suggesting their usefulness in breeding programs. While many published work are available on SSR primer design flows[124,125], only a few had combined bioinformatics with wet lab validations[126–128].

Current data suggests the necessity of such validations to capture the naturally existing biological variation, very common especially in the cross-pollinating or out-crossing species.

Conclusions

In this paper, we provided a methodology to be followed in assembling a transcriptome from raw data, as well as evaluate the accuracy of the assembly. In order to simplify the process and make it more comprehensible, we used freely available software and tools for the entire workflow. This allows the researcher to experiment and understand the flow of work without external challenges. While the presented methodology discusses the most popular tools used at each stage, it is recommended that the necessary tools are chosen according to the characteristics of the data as well as the end goal of the transcriptome assembly. Furthermore, we utilized the assembled transcriptome for one important application – identification of gene-specific SSR markers – for *S. album* breeding programs. All the designed markers amplified successfully, validating the designed workflow.

To best of our knowledge, this is the first validated attempt of a bioinformatics workflow for de novo transcriptome assembly followed by SSR primer design using freely available software. Most importantly, the bits and pieces of the process are connected in a user-friendly manner, facilitating efforts of biologists. Most of the available pipelines for transcriptome assembly are completely automated, and the work packages are bundled together. Other than a few environmental configurations, the user is not asked to manually examine or handle intermediate outputs during the assembly process in these pipelines. Rather, they are provided with the means to supply raw sequence data as input to the pipeline, and receive a complete or draft transcriptome assembly as the final output. A drawback of this fully-automated approach is that novel biologists may find it too ambiguous as to what happens during the assembly process. As a solution, our workflow is separated into individual modules that are executed separately. This allows the user the flexibility to observe and handle intermediate files, providing a greater depth of understanding as to what is happening with the data at each stage of the process. For beginner biologists, this would be very helpful in understanding the fundamentals of RNA-seq data and transcriptome assembly.

Also, having individual scripts for each of the stages means that the user could easily use different individual scripts simultaneously on different sets of data without affecting the outcome. For an example, the user can pre-process the dataset A using the relevant script for quality control, while at the same time running the transcriptome assembly script on dataset B that had been already pre-processed, even as they are designing primers for dataset C. This would allow users to work efficiently, while working on multiple analyses simultaneously. Having a bundled, automated end-to-end pipeline would prevent the user from using it on multiple datasets which are at different processing stages at once. Another advantage of the separated components is that the user can easily branch out or extend the workflow into other experiments by integrating new user-defined or already available tools. If analysis priorities were to change, it should be relatively easy to modify and re-direct the workflow. This would prove very useful in building an in-house RNA-seq assembly and analysis system for research teams and labs at no cost. In addition, since no complex set-up of the environment is expected of the user other than installing the necessary individual programs, anyone without great knowledge or background in computer science could easily use the scripts to analyze their data.

Therefore, it is evident that our workflow for de novo transcriptome assembly and SSR primer design is simple, comprehensive in dealing with necessary stages required to assemble the transcriptome and design SSR primers, yet complete in providing a workflow starting from raw RNA-seq data to analysis. Considering that it extends to primer design, it is a unique workflow among the cohort of transcriptome assembly pipelines. It is favorable for small institutions and research teams, as a solution for their RNA-seq analysis needs under very low budgets but greater research objectives.

Data availability

The RNA_seq dataset analysed during the current study is publicly available in the NCBI repository, deposited by a previously published paper (BioProject PRJNA297453), https://trace.ncbi.nlm.nih.gov/Traces/study/?acc=SRP064355. The gene data are also publicly available at accessions KC842188.1, KT160233.1, KT160234.1, KT160235.1 , KT160236.1, KT160237.1, KT160238.1, KT160239.1, NC_000932.1 (54,958.0.56397), and NC_003070.9 (28,451,138.0.28453820, complement) in the NCBI repository under previous original submissions.

Received: 8 December 2019; Accepted: 21 September 2020
Published online: 26 October 2020

References

1. Sboner, A., Mu, X., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome Biol.* **12**, 125 (2011).
2. Telenti, A. *et al.* Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci.* **113**, 11901–11906 (2016).
3. Salipante, S. J. *et al.* Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. *Genome Res.* **25**, 119–128 (2015).
4. Zimin, A. V. *et al.* Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* **27**, 787–792 (2017).
5. Wang, B. *et al.* Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* **7**, 11708 (2016).
6. Sun, Y. *et al.* Fish-T1K (Transcriptomes of 1,000 Fishes) Project: large-scale transcriptome data for fish evolution studies. *Gigascience* **5**, 18 (2016).
7. Vicentini, R. *et al.* Large-Scale Transcriptome Analysis of Two Sugarcane Genotypes Contrasting for Lignin Content. *PLoS ONE* **10**, e0134909 (2015).
8. Pareek, C. S., Smoczynski, R. & Tretyn, A. Sequencing technologies and genome sequencing. *J. Appl. Genet.* **52**, 413–435 (2011).

9. Metzker, M. L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11**, 31–46 (2010).
10. Pareek, C. S., Smoczynski, R. & Tretyn, A. Sequencing technologies and genome sequencing. J. Appl. Genet. **52**, 413–435 (2011).
11. Applied Biosystems Genetic Analysis Systems. https://www.thermofisher.com/lk/en/home/life-science/sequencing/sanger-sequencing/sanger-sequencing-technology-accessories.html.
12. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
13. Illumina. *An introduction to Next-Generation Sequencing Technology*. www.illumina.com/technology/next-generation-sequencing.html.
14. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
15. Wang, Y. & Navin, N. E. Advances and applications of single-cell sequencing technologies. *Mol. Cell* **58**, 598–609 (2015).
16. Thermo Fisher Launches New Systems to Focus on Plug and Play Targeted Sequencing|GenomeWeb. https://www.genomeweb.com/sequencing-technology/thermo-fisher-launches-new-systems-focus-plug-and-play-targeted-sequencing.
17. Ambardar, S., Gupta, R., Trakroo, D., Lal, R. & Vakhlu, J. High throughput sequencing: an overview of sequencing chemistry. *Indian J. Microbiol.* **56**, 394–404 (2016).
18. Rhoads, A. & Au, K. F. PacBio sequencing and its applications. *Genom. Proteomics Bioinform.s* **13**, 278–289 (2015).
19. Minio, A., Lin, J., Gaut, B. S. & Cantu, D. How single molecule real-time sequencing and haplotype phasing have enabled reference-grade diploid genome assembly of wine grapes. *Front. Plant Sci.* **8**, 826 (2017).
20. Branton, D. *et al.* The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* **26**, 1146–1153 (2008).
21. Kono, N. & Arakawa, K. Nanopore sequencing: review of potential applications in functional genomics. *Dev. Growth Differ.* **61**, 316–326 (2019).
22. PromethION. https://nanoporetech.com/products/promethion.
23. Longer and longer: DNA sequence of more than two million bases now achieved with nanopore sequencing. https://nanoporetech.com/about-us/news/longer-and-longer-dna-sequence-more-two-million-bases-now-achieved-nanopore.
24. Mendoza, E. A., Neumann, A., Kuznetsova, Y., Brueck, S. R. J. & Edwards, J. Electrophoretic plasmonic nanopore biochip genome sequencer. *Opt. Laser Technol.* **109**, 199–211 (2019).
25. Sequencing Platforms | Compare NGS platform applications & specifications. https://www.illumina.com/systems/sequencing-platforms.html.
26. Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucl. Acids Res.* **36**, e105–e105 (2008).
27. Scholz, M. B., Lo, C.-C. & Chain, P. S. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Curr. Opin. Biotechnol.* **23**, 9–15 (2012).
28. Schneeberger, K. *et al.* Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. *Proc. Natl. Acad. Sci.* **108**, 10249–10254 (2011).
29. Wang, B., Ekblom, R., Bunikis, I., Siitari, H. & Höglund, J. Whole genome sequencing of the black grouse (Tetrao tetrix): reference guided assembly suggests faster-Z and MHC evolution. *BMC Genomics* **15**, 180 (2014).
30. Garg, R., Patel, R. K., Tyagi, A. K. & Jain, M. De Novo assembly of Chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.* **18**, 53–63 (2011).
31. Wang, Z. *et al.* De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). *BMC Genomics* **11**, 726 (2010).
32. Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
33. Dong, X. *et al.* De novo assembly of a wild pear (*Pyrus betuleafolia*) genome. *Plant Biotechnol. J.* **18**, 581–595 (2020).
34. Daccord, N. *et al.* High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat. Genet.* **49**, 1099–1106 (2017).
35. Huang, J. *et al.* De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **6**, 1 (2017).
36. Nock, C. J. *et al.* Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnol. J.* **9**, 328–333 (2011).
37. Pop, M. Genome assembly reborn: recent computational challenges. *Brief. Bioinform.* **10**, 354–366 (2009).
38. Alkan, C., Sajjadian, S. & Eichler, E. E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
39. Schatz, M. C., Witkowski, J. & McCombie, W. R. Current challenges in de novo plant genome sequencing and assembly. *Genome Biol.* **13**, 243 (2012).
40. Medvedev, P., Georgiou, K., Myers, G. & Brudno, M. Computability of models for sequence assembly. In *Algorithms in Bioinformatics. WABI 2007. Lecture Notes in Computer Science* (eds Giancarlo, R. & Hannenhalli, S.), vol. 4645 LNBI 289–301 (2007).
41. Reinhardt, J. A. *et al.* De novo assembly using low-coverage short read sequence data from the rice pathogen *Pseudomonas syringae* pv. oryzae. *Genome Res.* **19**, 294–305 (2008).
42. Page, A. J. *et al.* Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb. Genomics* **2**, e000083 (2016).
43. Ouborg, N. J., Piquot, Y. & Van Groenendael, J. M. Population genetics, molecular markers and the study of dispersal in plants. *J. Ecol.* **87**, 551–568 (1999).
44. Semagn, K., Bjørnstad, Å. & Ndjiondjop, M. N. An overview of molecular marker methods for plants. *Afr. J. Biotechnol.* **5**, 2540–2568 (2006).
45. Mohan, M. *et al.* Genome mapping, molecular markers and marker-assisted selection in crop plants. *Mol. Breed.* **3**, 87–103 (1997).
46. Grover, A. & Sharma, P. C. Development and use of molecular markers: past and present. *Crit. Rev. Biotechnol. Dev. Use Mol. Markers Past Present*. https://doi.org/10.3109/07388551.2014.959891 (2014).
47. Nadeem, M. A. *et al.* DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnol. Biotechnol. Equip.* **32**, 261–285 (2018).
48. Gupta, P. K., Balyan, H. S., Sharma, P. C. & Ramesh, B. Microsatellites in plants: anew class of molecular markers. *Curr. Sci.* **70**, 45–54 (1996).
49. Liang, X. *et al.* Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and Arachis wild species. *BMC Plant Biol.* **9**, 35 (2009).
50. Triwitayakorn, K. *et al.* Transcriptome Sequencing of *Hevea brasiliensis* for development of microsatellite markers and construction of a genetic linkage map. *DNA Res.* **18**, 471–482 (2011).
51. Harmon, M. *et al.* Development of novel genic microsatellite markers from transcriptome sequencing in sugar maple (Acer saccharum Marsh.). *BMC Res. Notes* **10**, 1–7 (2017).
52. Lu, Q.-X. *et al.* Development of 19 novel microsatellite markers of lily-of-the-valley (Convallaria, Asparagaceae) from transcriptome sequencing. **47**, 3041–3047 (2020).
53. El-Metwally, S., Hamza, T., Zakaria, M. & Helmy, M. Next-generation sequence assembly: four stages of data processing and computational challenges. *PLoS Comput. Biol.* **9**, e1003345 (2013).
54. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
55. Mundry, M., Bornberg-Bauer, E., Sammeth, M. & Feulner, P. G. D. Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach. *PLoS ONE* **7**, e31410 (2012).

56. Bradnam, K. R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**, 10 (2013).
57. Smith, D. R. Buying in to bioinformatics: an introduction to commercial sequence analysis software. *Brief. Bioinform.* **16**, 700–709 (2015).
58. Goecks, J. *et al.* Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, (2010).
59. Langmead, B. & Nellore, A. Cloud computing for genomic data analysis and collaboration. *Nat. Rev. Genet.* **19**, 208–219 (2018).
60. Amazon EC2 Pricing - Amazon Web Services. https://aws.amazon.com/ec2/pricing/.
61. Wolstencroft, K. *et al.* The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Res.* **41**, (2013).
62. Arvados|Open Source Big Data Processing and Bioinformatics. https://arvados.org/.
63. Dooley, R., Vaughn, M., Stanzione, D., Terry, S. & Skidmore, E. Software-as-a-Service: The iPlant Foundation AP. https://foundation.iplantcollaborative.org.
64. D'Antonio, M. *et al.* RAP: RNA-Seq analysis pipeline, a new cloud-based NGS web application. *BMC Genomics* **16**, (2015).
65. Sadedin, S. P., Pope, B. & Oshlack, A. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics* **28**, 1525–1526 (2012).
66. Nextflow - A DSL for parallel and scalable computational pipelines. https://www.nextflow.io/.
67. Seoane, P. *et al.* TransFlow: amodular framework for assembling and assessing accurate de novo transcriptomes in non-model organisms. *BMC Bioinform.* **19**, (2018).
68. Vitturi, R., Colomba, M., Pirrone, A. & Mandrioli, M. WGSSAT: A high-throughput computational pipeline for mining and annotation of SSR markers from whole genomes. *J. Hered.* **93**, 279–282 (2002).
69. Mokhtar, M. M. & Atia, M. A. M. SSRome: an integrated database and pipelines for exploring microsatellites in all organisms. *Nucleic Acids Res.* **47**, D244–D252 (2019).
70. Celedon, J. M. *et al.* Heartwood-specific transcriptome and metabolite signatures of tropical sandalwood (*Santalum album*) reveal the final step of (Z)-santalol fragrance biosynthesis. *Plant J.* **86**, 289–299 (2016).
71. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
72. SRA and other NCBI databases. https://www.ncbi.nlm.nih.gov/sra/docs/#sra-and-other-ncbi-databases.
73. SRA database growth. https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/.
74. SRA Toolkit download. https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/.
75. NCBI FTP Directory. ftp://ftp.ncbi.nih.gov/.
76. Andrews, S., FastQC. Babraham Bioinformatics. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
77. FASTX-Toolkit. https://hannonlab.cshl.edu/fastx_toolkit/index.html.
78. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
79. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
80. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
81. Honaas, L. A. *et al.* Selecting superior de novo transcriptome assemblies: lessons learned by leveraging the best plant genome. *PLoS ONE* **11**, e0146062 (2016).
82. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
83. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
84. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
85. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* **26**, 1134–1144 (2016).
86. Diaz-Chavez, M. L. *et al.* Biosynthesis of sandalwood oil: *Santalum album* CYP76F cytochromes P450 produce santalols and bergamotol. *PLoS ONE* **8**, e75053 (2013).
87. You, F. M. *et al.* BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinform.* **9**, 253 (2008).
88. OligoAnalyzer Tool - primer analysis|IDT. https://sg.idtdna.com/pages/tools/oligoanalyzer.
89. Healey, A., Furtado, A., Cooper, T. & Henry, R. J. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant Methods* **10**, 21 (2014).
90. Rabah, S. O. *et al.* Plastome sequencing of ten nonmodel crop species uncovers a large insertion of mitochondrial DNA in cashew. *Plant Genome* **10**, 0 (2017).
91. Nie, X. *et al.* Complete chloroplast genome sequence of a major invasive species, crofton weed (*Ageratina adenophora*). *PLoS ONE* **7**, e36869 (2012).
92. Wu, Z. *et al.* A precise chloroplast genome of *Nelumbo nucifera* (Nelumbonaceae) evaluated with Sanger, Illumina MiSeq, and PacBio RS II sequencing platforms: insight into the plastid evolution of basal eudicots. *BMC Plant Biol.* **14**, 289 (2014).
93. Visualizing size and GC content of genomes|Kaggle. https://www.kaggle.com/camnugent/visualizing-size-and-gc-content-of-genomes.
94. Mullis, K. *et al.* Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* **51**, 263–273 (1986).
95. Abd-Elsalam, K. A. Bioinformatic tools and guideline for PCR primer design. *Afr. J. Biotechnol.* **2**, 91–95 (2003).
96. Alhakami, H., Mirebrahim, H. & Lonardi, S. A comparative evaluation of genome assembly reconciliation tools. *Genome Biol.* **18**, 93 (2017).
97. Dapas, M., Kandpal, M., Bi, Y. & Davuluri, R. V. Comparative evaluation of isoform-level gene expression estimation algorithms for RNA-seq and exon-array platforms. *Brief. Bioinform.* **18**, bbw016 (2016).
98. Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nat. Rev. Genet.* **14**, (2013).
99. Earl, D. *et al.* Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.* **21**, 2224–2241 (2011).
100. Salzberg, S. L. *et al.* GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22**, 557–567 (2012).
101. Hunt, M., Newbold, C., Berriman, M. & Otto, T. D. A comprehensive evaluation of assembly scaffolding tools. *Genome Biol.* **15**, R42 (2014).
102. Transcriptome Assembly Quality Assessment · trinityrnaseq/trinityrnaseq Wiki · GitHub. https://github.com/trinityrnaseq/trinityrnaseq/wiki/Transcriptome-Assembly-Quality-Assessment.
103. Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **43**, e37–e37 (2015).
104. Abnizova, I., te Boekhorst, R. & Orlov, Y. L. Computational errors and biases in short read next generation sequencing. *J. Proteomics Bioinform.* **10**, 1–17 (2017).

105. Ma, X. *et al.* Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* **20**, 50 (2019).
106. Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* **5**, 1005–1010 (2008).
107. Guo, Y., Ye, F., Sheng, Q., Clark, T. & Samuels, D. C. Three-stage quality control strategies for DNA re-sequencing data. *Brief. Bioinform.* **15**, 879–889 (2014).
108. Xie, Y. *et al.* SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660–1666 (2014).
109. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086–1092 (2012).
110. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909–912 (2010).
111. Tao, X. *et al.* Digital gene expression analysis based on integrated de novo transcriptome assembly of sweet potato [*Ipomoea batatas* (L.) Lam.]. *PLoS ONE* **7**, e36234 (2012).
112. Liu, S., Li, W., Wu, Y., Chen, C. & Lei, J. D. Novo transcriptome assembly in Chili Pepper (*Capsicum frutescens*) to identify genes involved in the biosynthesis of capsaicinoids. *PLoS ONE* **8**, e48156 (2013).
113. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* **95**, 315–327 (2010).
114. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
115. Trinity's In silico Read Normalization. https://github.com/trinityrnaseq/trinityrnaseq/wiki/Trinity-Insilico-Normalization.
116. Durai, D. A. & Schulz, M. H. In silico read normalization using set multi-cover optimization. *Bioinformatics* **34**, 3273–3280 (2018).
117. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
118. Crusoe, M. R. *et al.* The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Research* **4**, 900 (2015).
119. Lowe, E. K., Swalla, B. J. & Titus Brown, C. Evaluating a lightweight transcriptome assembly pipeline on two closely related ascidian species. *PeerJ Prepr.* https://doi.org/10.7287/peerj.preprints.505v1 (2014).
120. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **12**, 671–682 (2011).
121. Iorizzo, M. *et al.* A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.* **48**, 657–666 (2016).
122. Baker, M. D. novo genome assembly: what every biologist should know. *Nat. Methods* **9**, 333–337 (2012).
123. Phillippy, A. M., Schatz, M. C. & Pop, M. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.* **9**, R55 (2008).
124. Mccouch, S. R. *et al.* Development and Mapping of 2240 New SSR Markers for Rice (Oryza sativa L.). *DNA Research* vol. 9 (2002).
125. Zalapa, J. E. *et al.* Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Am. J. Bot.* **99**, 193–208 (2012).
126. Kaur, S. *et al.* Transcriptome sequencing of lentil based on second-generation technology permits large-scale unigene assembly and SSR marker discovery. *BMC Genomics* **12**, 265 (2011).
127. Parchman, T. L., Geist, K. S., Grahnen, J. A., Benkman, C. W. & Buerkle, C. A. Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC Genomics* **11**, 180 (2010).
128. Wang, H. *et al.* Next-generation sequencing of the *Chrysanthemum nankingense* (Asteraceae) transcriptome Permits large-scale unigene assembly and SSR marker discovery. *PLoS ONE* **8**, e62293 (2013).

Acknowledgements

The authors would like to thank the staff of the Agricultural Biotechnology Centre, Faculty of Agriculture, University of Peradeniya for the continuous support given. The authors thank Dr. Ardeshir B. Damania, Department of Plant Science, University of California Davis, USA for helpful editorial comments provided for improving the manuscript.

Author contributions

D.N.U. performed designing of the work, workflow development, microsatellite markers development and a major contributor in writing the manuscript. C.H.W.M.R.B. performed the lab validation of microsatellite markers and writing the manuscript. P.C.G.B. provided the funding and guidance, and revised the manuscript. A.U.B. helped with designing and improving the work, major supervision and revised the manuscript. All authors read and approved the final manuscript.

Funding

An institutional grant from National Science Foundation (NSF), Sri Lanka under the initiatives of Department of Small Industries, Sri Lanka provided the stipend for the Research Assistant position of DNU and her graduate studies (M.Phil.) fees. Sri Lanka Council for Agricultural Policy (CARP) provided funds for materials, equipment, and bench work undertaken by CHWMRB in contribution to this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-75270-8.

Correspondence and requests for materials should be addressed to A.U.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.