Application Note

De novo assembly of paired-end plant transcriptome data

Application note on *de novo* assembly of a plant transcriptome using Illumina paired-end sequencing data from *Lactuca serriola* (prickly lettuce) in CLC Genomics Workbench

De novo assembly of paired-end plant transcriptome data

CLC bio's de novo assembly tool allows for initiation of genomic data analysis in organisms without previously sequenced genomes. This is especially important in plant and animal genome research where the majority of species do not have reference sequences available. The de novo assembler was tested over the years by many plant and animal researchers who demonstrated that it successfully assembled short sequencing reads into long contigs. This application note provides some guidelines for de novo assembly of transcriptome high-throughput sequencing reads using CLC Genomics Workbench.

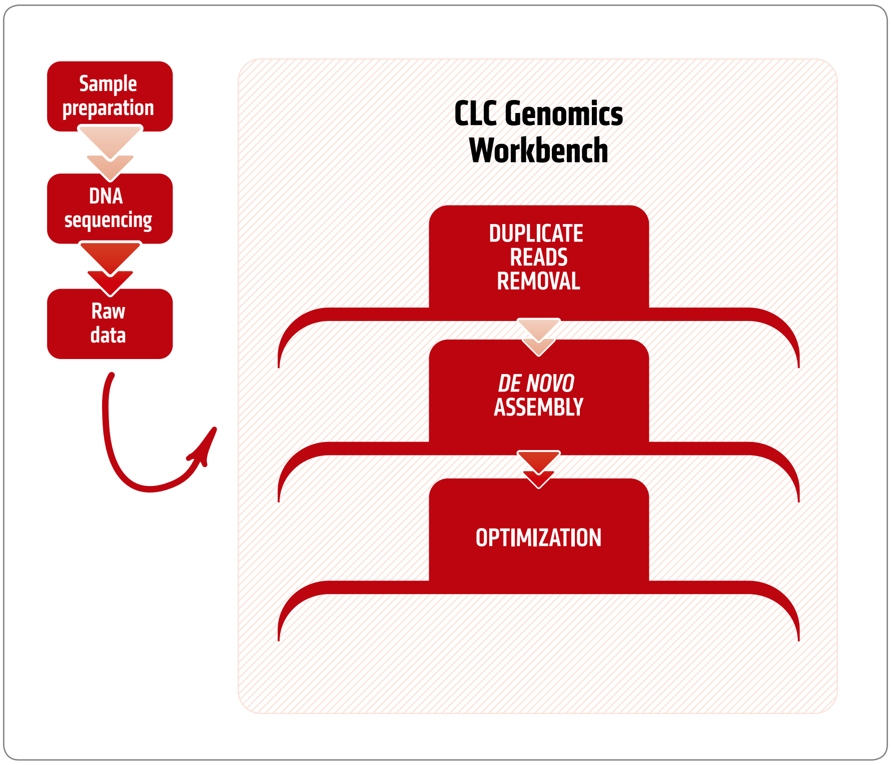

Figure 1: Work flow for *de novo* assembly using CLC Genomics Workbench

Reads dataset

This application note is using the publically available reads from *Lactuca serriola* (prickly lettuce) mRNA library[1,2]. The library was prepared using the Illumina mRNA-Seq library protocol followed by DSN (Duplex Specific Nuclease) normalization. The data set consists of around 93 Million (46.5M X 2) paired end reads (85 nt) from Illumina Genome Analyzer IIx.

Figure 2: *Lactuca serriola*

Analysis workflow

Duplicate Reads Removal

Removal of duplicated reads prior to *de novo* assembly may be important for some data sets. This applies especially to *de novo* assembly of reads produced from standard (non-normalized) mRNA libraries.

The duplicate reads removal tool of CLC Genomics Workbench filters out duplicate reads which have been created during PCR amplification and which can have a distorting influence on assembly since some sequences will be present in artificially high numbers (Fig. 3). Also, PCR amplification of heavily expressed transcripts can lead to read duplication.

The data set used in this note was produced from a normalized mRNA library. However, we were still able to eliminate more than 11 million duplicated reads with the duplicate reads removal tool.

De novo Assembly

The *de novo* assembly algorithm of CLC Genomics Workbench supports hybrid data, meaning you can perform assemblies that use both short and long reads, paired reads as well as reads from different sequencing technologies: all these can be entered into the same assembly job.

The *de novo* assembly algorithm is divided into two phases. In the first phase, simple contig sequences are created by using all the information that is in the read sequences. This is the actual *de novo* assembly part of the process, however, these contig sequences do not contain any information about which reads were used to create the contigs. In the second phase, all reads are mapped back to the simple contig sequences. This is done to show e.g. coverage levels along the contigs and enables more downstream

[1] mRNA library submitted to GenBank SRA (Short Read Archive) by University of California – Davis researchers. Compositae Genomics Project http://compgenomics.ucdavis.edu

[2] Data is available at http://www.ncbi.nlm.nih.gov/sra/SRX098217

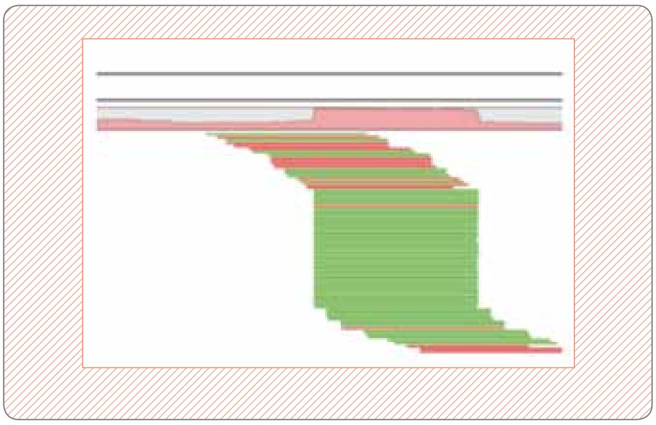

Figure 3: Stacks of duplicated reads amongst the mapped reads visualized in mapping view in CLC Genomics Workbench.

analysis like variant detection. When starting the assembly you can choose whether you want to output simple contigs only, or have the reads mapped back, too. For reporting, you can choose a summary report on the assembly and a sequence list of the unmapped reads.

Parameters have to be specified for each set of reads. Paired information is used for both the first and the second phase of the assembly but the other parameters relate only to the second phase where the reads are mapped back to the contig sequences.

The parameter settings include penalty scores for mismatch, insertion and deletion costs, length fraction and similarity. For the data set used here, we selected the heaviest penalty for mismatch, insertion, and deletion costs which is 3. Length fraction was set to 0.5 and similarity was set to 0.95. This means that at least 50% of the individual read needs to have at least 95% identity with the contig sequence in order to be included in the mapping.

You can also specify the minimum contig length. We set the minimum contig length to 300 nt so contigs shorter than this will not be reported.

Word size	24	42	48	54	60	64
% reads matched	95.4	95.3	94.8	94.1	93.9	92.3
# contigs, x1000	66.9	54.8	50.7	46.5	42.5	39.9
Average contig length	919	954	959	966	959	940
Assembly length in Mb	61.5	52.3	48.7	44.9	40.7	37.5

Table 1: Effect of word size on assembly outcome

Furthermore, you can specify how to deal with non-specific reads and reporting of conflicts.

Finally, our *de novo* assembler allows the user to either manually set the word size or it can be automatically chosen by the algorithm. For the first *de novo* assembly run we selected to use the automatic word size.

Results and optimization

The assembler selected the word size of 24. The input was around 81 million reads and the number of resulting contigs was 66.9K with an average contig length of 919 nt.

Optimization of word size

To optimize the assembly output you can change the parameters, e.g. the word size.

In our experiment, we repeated the *de novo* assembly on the same data set

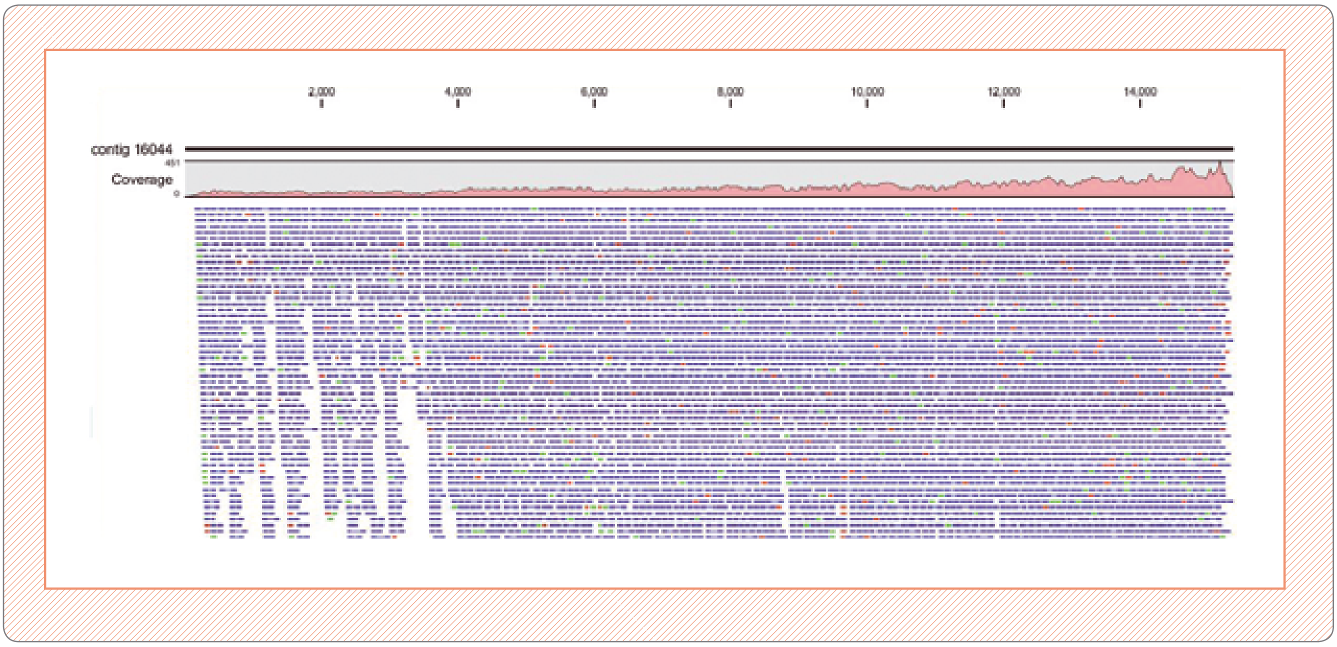

Figure 4: The longest assembled contig with 15,368 nt in the read mapping view of CLC Genomics Workbench. This contig represents the longest plant transcript, a homolog of the *Arabidopsis* auxin transport protein (BIG) gene.

with different word sizes as listed in Table 1. The table shows that the word size of 54 produced the longest (on average) contigs and a graphical representation is shown in Figure 5.

Performing a *de novo* assembly with word size 54, we obtained 46.5K contigs with the final assembly length 45 Mb. The average size of the contigs was 966 nt and in total 94.1% of the reads were mapped.

It is difficult to provide any general recommendations on the optimal word size but for high quality reads the word size should generally be higher than for low quality reads.

To further improve the assembly, other parameters such as the alignment stringency can also be adjusted.

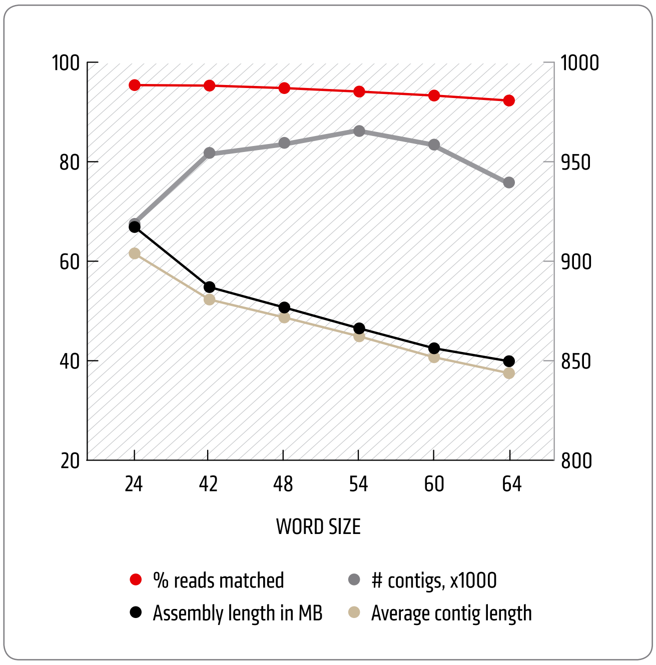

Figure. 5: The results from six assemblies of *Lactuca serriola* reads using different word sizes

8.10.2012 CLC bio