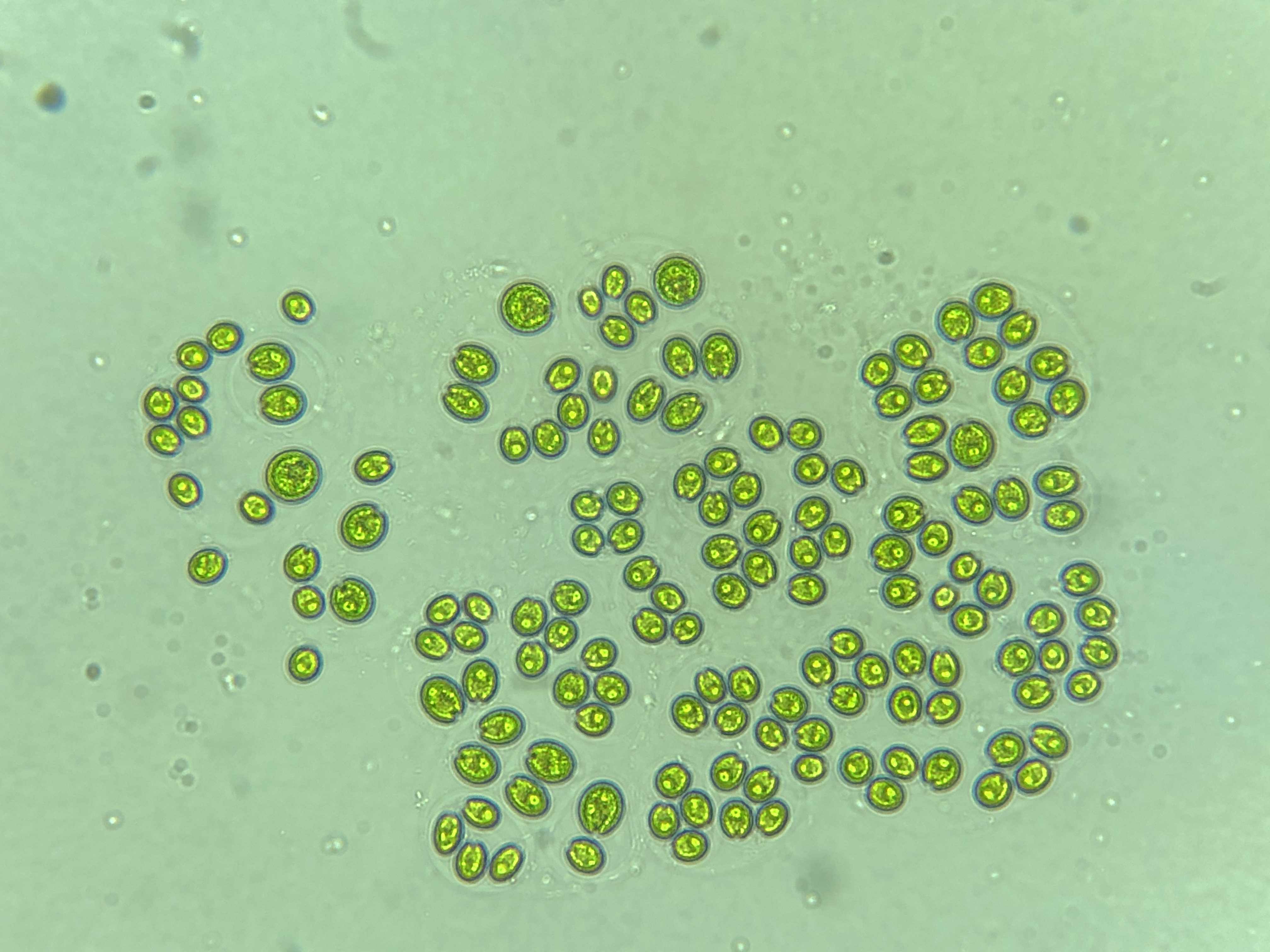
**Gene expression in unicellular, palmelloid and multicellular states of *Chloromonas oogama* – Report of the pilot study**

**Background**

Three states exist in the unicellular green algae, *Chloromonas oogama*, as shown in the photo below:



Unicellular

Palmelloid

Multicellular

Pyrenoid?

Figure 1. Three cell states in *Chloromonas oogama*.

*C. oogama* is a species of green algae closely related to *Chlamydomonas reinhardtii*. It is written in several places that this species lacks a structure called pyrenoid. In eukaryotic micro-algae, the Rubisco micro-compartment, which catalyzes the first major step of carbon fixation, is called pyrenoid. In our conversation, Maria mentioned that *C. oogama* used in this study has pyrenoid visible in the photos. Is this species then really *C. oogama*?

The three cell states in *C. oogama* are the unicellular, palmelloid (consisting of 2-4 cells) and a multicellular structure. We do not have information on *C. oogama* cell cycle but we can learn a bit from the cell cycle in *Chlamydomonas. Chlamydomonas* proliferates using a modified cell cycle, termed multiple fission (Figure 2). Multiple fission cell cycles are characterized by a prolonged growth phase (G1), during which cells can enlarge by more than two-fold in size. At the end of G1, *Chlamydomonas* cells undergo successive rounds of rapidly alternating S phases and mitoses (S/M) to produce 2n daughter cells. Daughters then hatch out of the mother cell to begin the cycle again. The number of S/M cycles that each mother cell undergoes is dictated by cell size (Craigie and Cavalier-Smither 1982; Donnan and John, 1983). Depending on growth conditions a mother cell undergoes between one and five S/M cycles to produce 2, 4, 8, 16 or 32 daughters (Lien and Knutsen, 1979). Under a typical diurnal cycle (e.g. 12 h of light/12 h of dark) the cell cycle becomes synchronized such that growth occurs during the light phase and cell division (S/M) occurs in the dark. There are many more details about the cell cycle that we can think about, for example, when is the commitment point in G1 and how is the number of cell divisions determined in the cell? As we learn more about cell cycles in green algae, we can add more details. However, it is useful to list four categories of genes that are vital to cell cycle regulation in *Chlamydomonas* and other eukaryotes. These include **Cyclin-dependent kinases**, **Cyclins**, **Retinoblastoma (RB) tumor-suppressor pathway** and **Anaphase promoting complex/cyclosome (APC/C)**. When one reads gene names as such, they mean nothing. As we progress, we will learn more about the molecular mechanism and biochemistry of each protein involved so little by little, names will become meaningful. For now, let’s move on to our study of gene expression. (The information on cell cycle is from Cross and Umen, 2015)

Diagram

Description automatically generated

Figure 2. *Chlamydomonas* multiple-fission cell cycle (Cross and Umen, 2015).

**Materials and Methods**

**Cell culture and sorting**

Mixed samples were sorted with the Copas Vision™ (Union Biometrica, Holliston, MA) by size. Regions in a scatterplot of Time of flight (TOF, equivalent to length of particle) and extinction (equivalent to density) were delineated to include the desired phenotypes. Three phenotypes were delineated: single cells, four-celled palmelloids, and clusters with at least ca. 20 cells. We aimed to sort 10 000 particles per phenotype. This worked well for single cells and palmelloids, but fewer clusters were sorted due to time constraints. If we estimate 20 cells per cluster (minimum), all cluster samples included more than 10 000 individual cells. Particles were deposited into 50 mL Falcon tubes and cells were harvested by centrifuging these tubes and removing the supernatant. RNA was extracted with the RNeasy Micro plant kit from Qiagen (Table 1).

Table 1. Concentration, volume and RIN (RNA Integrity Number).

|  |  |  |  |
| --- | --- | --- | --- |
| **SAMPLE ID** | **CONCENTRATION** | **VOLUME** | **RIN** |
| Osg03Sin | 4.92 | 25 | 5.4 |
| Osg03Pal | 5.00 | 25 | 5.8 |
| Osg03Mul | 2.2 | 25 | 4.4 |
| Osg10Sin | 3.84 | 25 | 4.7 |
| Osg10Pal | 0.02 | 25 | 4.1 |
| Osg10Mul | 2.92 | 25 | 5.7 |
| Osg11Sin | 1.32 | 25 | 6.7 |
| Osg11Pal | 0.02 | 25 | unknown |
| Osg11Mul | 0.24 | 25 | 4.1 |
| Osg12Sin | 7.28 | 25 | 7 |
| Osg12Pal | 0.88 | 25 | 2.4 |
| Osg12Mul | 0.24 | 25 | 3.8 |

**Library preparation and sequencing**

Sequencing libraries were prepared using Illumina Stranded Total RNA Prep, Ligation with Ribo-zero plus. Cluster generation and paired-end sequencing was done in one NovaSeq 6000 SP lane with read length of 150 bases (insert size = ?). There are two important aspects of library preparation to consider. The first is strandedness and the second is the method used for rRNA depletion. Let’s understand the term “stranded” here because it is very important and useful when one is trying to assemble RNA-seq reads and I am always confused about it. Figure 3 shows the steps taken in preparation of our stranded library. The aim is to preserve the strand information from which the RNA was transcribed. The dUTP method was used in our case and is the most accurate method and is shown below. Strand-specificity is achieved by the second-strand cDNA incorporating dUTPs instead of dTTPs. Digestion of dUTPs by uracil-DNA glycosylase (UGDase) prevents this strand from being PCR amplified, conferring single-strand specificity (Martin et al, 2013).

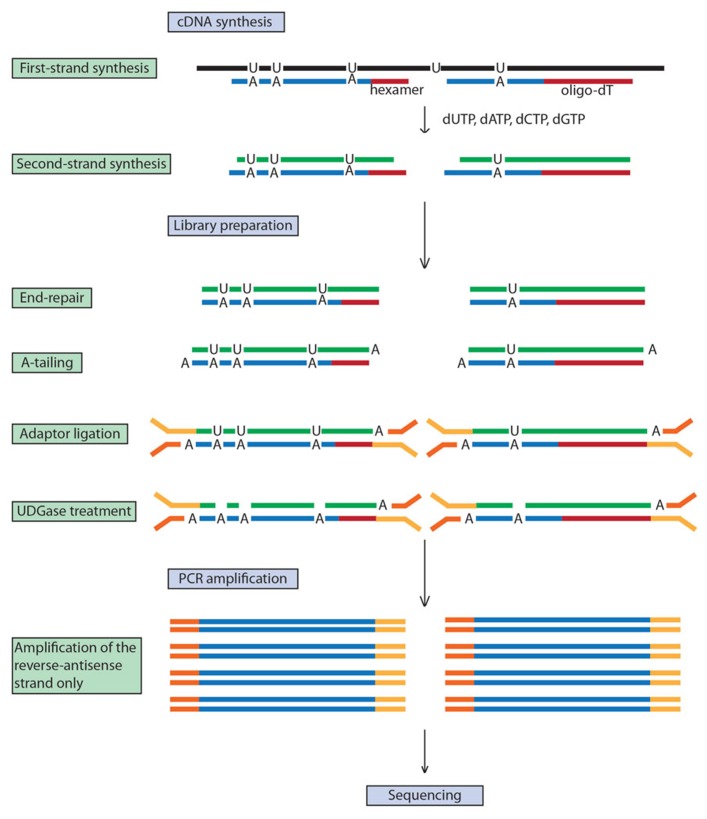


Figure 3. Steps in producing stranded library.

Now let’s think about the way in our library preparation we have gotten rid of the rRNA molecules. There are two ways for it. It is possible to enrich for transcripts with Poly-A to select only mRNA molecules or to deplete the rRNA molecules. When sample concentration and quality is low, the second approach is chosen. The problem with Ribo-zero plus approach is that many reads might originate from intronic or intergenic regions. In a study comparing different library preparation methods, they showed that to achieve the same gene detection level, approximately 4 times more reads are necessary when sequencing with Ribo-zero plus compared to Poly-A method (Zhao et al, 2014, BMC Genomics). After adapter and contamination removal, almost half of the reads across all samples were rRNA. The question is why such large number of rRNAs are remained in the sample? In the rRNA depletion method, some probes are hybridized to the rRNA molecule. Could it be that the probes do not attach well to the rRNA of our algae?

The conclusion is that while choosing a stranded protocol is a great choice, by obtaining more RNA sample, choosing the Poly-A library method where mRNAs are enriched could increase the number of reads obtained from genic regions. This means with fewer reads obtained from Poly-A mRNA enrichment method, we can have more power to detect expression. If choosing RNA-depletion method, a larger number of reads must be generated.

**Raw read statistics, quality control and filtering**

I filtered the data in several steps. I should admit I made many mistakes, coming back every time to see how to correct it. I first removed all the Nextera adaptors using trimmomatic which removed between (X and X%) percentage of data. I then used Kraken2 against a bacterial, viral, plasmid and human database to remove contamination from these sources. To remove rRNA, I mapped all reads against all the ribosomal sequences (ribosomal 28S, 26S, 18S, 5.8S and the ITS sequence and chloroplast 16S, 23S and 5S) of C. reinhardtii. I extracted all paired reads where none of the mates map to the ribosomal sequences. This removed from (X – X) percent of data. I finally removed transcripts derived from chloroplast by mapping against the chloroplast genome of C. reinhardtii as they were marked by as over-represented sequences in some samples.

Table 2. Number of read pairs

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample | Raw read pairs | Read pairs after adaptor removal | Read pairs after contamination removal | Read pairs after rRNA removal | Read pairs after cpDNA removal |
| Osg10Sin | 38613164 | 16060016 | 11830432 | 7158367 | 6456338 |
| Osg12Sin | 40463366 | 15839946 | 15673264 | 8385230 | 6958634 |
| Osg03Sin | 28258686 | 11526439 | 11375740 | 5749367 | 4816128 |
| Osg11Sin | 24226009 | 8882048 | 8735042 | 4538880 | 3881399 |
| Osg10Pal | 30611959 | 12894102 | 11556507 | 6122303 | 5330315 |
| Osg12Pal | 29907662 | 8972217 | 8865670 | 4700984 | 3840004 |
| Osg03Pal | 37065999 | 7022738 | 6846830 | 3710347 | 3217315 |
| Osg11Pal | 30646894 | 13266828 | 13038745 | 6724388 | 5547860 |
| Osg10Mul | 22603175 | 6324156 | 4594204 | 2560102 | 2236171 |
| Osg12Mul | 29256523 | 11983847 | 11920185 | 6040694 | 4971296 |
| Osg03Mul | 25248234 | 11479498 | 11218051 | 5741596 | 4878319 |
| Osg11Mul | 29848148 | 15025749 | 14975812 | 7692642 | 6298582 |

I checked data quality using FastQC/0.11.9. Several FASTQC modules failed across samples. These include 1. per base sequence content, 2. per sequence GC content, 3. sequence duplication levels, 4. overrepresented sequences and 5. adapter content. An example of quality control from our data and one from a high quality RNA-seq from other sources is shown in Figure 4. Let’s go through each module and see if we understand why they are failing with our data:

1. Per base sequence content: This plot reports the percent of bases called for each of the four nucleotide positions. What is normal? The fluctuations in base composition in the first ~12 bases are normal if the library was produced by priming using random hexamers which is true for our case. This bias provides enrichment of a number of different K-mers at the 5’ end of the reads. This does not create problem in downstream analyses and does not require trimming. What is abnormal? Across the rest of the reads, we would expect to see the four bases in equal proportions with A%=T% and C%=G%. We observe lots of fluctuations which removed to a great extent after removal of rRNA (Figure 4).
2. Per sequence GC content: This plot gives a theoretical distribution of GC assuming uniform GC content for all reads. Figure 4 shows three main peaks, one centered around 65, one around 57 and the last around 47. Removal of adapters and in particular, rRNA and transcripts derived from chloroplast improves this graph significantly, however, reads with GC content below 50% remains.
3. Sequence duplication level: The blue line is percentage of a given sequence in the file which are present a given number of times. In Figure 4 we see that there is a larger number of transcripts which are present in more than 5000 or even 10000 copies in the sample. Removal of rRNA data greatly improved this graph.
4. Adapter content: As shown in Figure 4, there was a large proportion of reads that were just adapters. Removing Nextera paired-end adapters solved this issue.

Calendar

Description automatically generated

Figure 4. An example of quality graphs after each round of filtering. The four failed quality metrics are shown here. Per base sequence content improves significantly after removal of rRNA and resembles a good RNA library. Per sequence GC content improves significantly after the removal of rRNA and further removal of over-represented chloroplast RNAs. There are still however three smaller peaks around 45, 40 and 30 percent GC content. Sequence duplication level improves significantly after the removal of rRNA. Nextera-paired end adapters are removed.

**Transcriptome assembly**

Using Trinity/2.13.2, I made a transcriptome assembly with the filtered reads. Now some interesting numbers. Before running Kraken2 when all microbial samples were present, I obtained an assembly of 294175 contigs. After removing the contamination, the number of contigs reduced to 187760. This means that more than 100,000 contigs were assembled bacterial reads. Finally, after removing rRNA and chloroplast reads, the number of contigs were reduced only slightly to 187230 (Table 3). To assess assembly completeness, I used BUSCO. 1519 BUSCO groups were searched, 85.2 % complete BUSCOs were recovered from the assembly (1295: Complete BUSCOs, 632: Complete and single-copy BUSCOs, 663: Complete and duplicated BUSCOs). Another measure of assembly quality is the percentage of properly mapped reads. Properly mapped means 1. where both members of the pair are aligned, 2. in the correct orientation, 3. on the same contig, 4. without overlapping either end of the contig. Ideally this must be 100% if the contigs are properly assembled (Table 4). Another quality measure is the assessment of the proportion of contigs that have a match with the protein set of a closely related species. In our case, I used the protein set of *C. reinhardtii* (Table 5).

Table 3. Transcriptome assembly metrics

|  |  |
| --- | --- |
| Contig metrics |  |
| Number of sequences | 187230 |
| Smallest contig | 188 |
| Largest contig | 20577 |
| Number of bases | 127884013 |
| Mean length | 683.03 |
| Number with open reading frame (ORF) | 65154 |
| Mean ORF percent | 75.2 |
| GC | 0.65 |
| N50 | 1075 |

Table 4. Read mapping metrics of transcriptome assembly

|  |  |
| --- | --- |
| Contig metrics |  |
| Fragments | 58432361 |
| Fragments mapped | 21711671 |
| Proportion of good mapping | 0.32 |
| Proportion of good contigs | 0.8 |

Table 5. Comparative metrics of the transcriptomic assembly with *C. reinhardtii* proteome

|  |  |
| --- | --- |
| Contig metrics |  |
| Conditional Reciprocal Best Blast hits | 29469 |
| Number of contigs with CRBB | 29161 |
| Proportion of contigs with CRBB | 0.16 |
| Number of reference proteins with CRBB | 11070 |
| Proportion of reference proteins with CRBB | 0.57 |
| Cov25 (proportion Cov25) | 8221 (0.42) |
| Cov50 (proportion Cov50) | 6070 (0.31) |
| Cov75 (proportion Cov75) | 4363 (0.22) |
| Cov85 (proportion Cov85) | 3636 (0.19) |
| Cov95 (proportion Cov95) | 2531 (0.13) |
| Reference coverage | 0.24 |

**Transcript quantification**