Contents

1.0 **Cecile 454 Data Analysis:** Summary

- We have analyzed Roche 454 sequence data derived from the mitochondrial genome sequence and assembled it using a GenBank reference sequence (AJ973190) derived from *Glomus intraradices*, a species of arbuscular mycorrhizal fungus.

  In silico analysis of 255,966 raw reads was done using two approaches: reference sequence mapping to mtLSU rRNA (AJ973190) and *de novo* sequence assembly. The mtLSU rRNA (AJ973190: 1..40,1097..1459,1861..2607) reference sequence used in this analysis is 2,607 bps in length and is interrupted by two introns located at position: 41..1096 and 1460..1860.

- Quality assessment and quality control of the raw reads: quality control involved adapter trimming preceded by conversion of the .sff file to FASTQ (positional quality scores of the sequenced reads) format.

  Quality control involved trimming of the multiplex identifier or MID (a short barcode sequence used to label samples when multiplexing) and the linker sequence (used to link the amplicon primer together with the MID sequence) to allow efficient mapping to the reference sequence. The combined length of the MID and linker was 36 bps and can cause problems in mapping.

- Assembly/Mapping of the amplicons: de novo assembly is a pre-requisite to create longer contiguous sequences out of the relatively short reads that are on average 350-400 bps long. Several algorithms have been developed for mapping short read data such as illumina and ABI SOLiD data but they are not useful for handling 454/Roche reads. In this analysis, we used both open source MIRA (Mimicking Intelligent Read Assembly) Chevreux et al. (2004) and CLC Genomics Workbench, a commercial package, to assemble the amplicons.

  We successfully mapped 58.5% (149,654) of the sequenced reads to the reference sequence out of which 75,754 reads were in forward orientation and 73,900 reads were in the reverse orientation.
  The remaining **41.5%** (106,312) of the reads did not map to the reference sequence and were assembled using a *de novo* approach. *De novo* assembly of the unmapped reads generated 248 contigs with an N50 of 418 (computed to estimate the quality of the assembly) and the longest contig was 1,219 base pairs (Table 2.24). To investigate the features of the contigs, we searched (BLAST) them against public databases.

  Coverage per amplicon and per MID was heterogeneous and between 0 and 15,264 (mean 7,276.404) for pool one, 2 and 10,217 (mean 3,211.616) for pool two, and between 0 and 10,141 (mean 3,925.497) for pool three. This heterogeneity (or 'spread factor') can be attributed to differences in PCR efficiency and suboptimal pooling of samples.

  The total coverage of the reference sequence by the reads was 99.9% while the no coverage (zero-coverage) regions were 2 (see tables 2.11, 2.16 and 2.22). One of the zero-coverage regions (2..999) overlaps the intron position (41..1096) of the mtLSU rRNA gene.
  The attached file (454Reads.MID_trimmed_coverage_of the_refseq.pdf) depicts regions of the short reads that are deleted and those that are highly (amplified) covered relative to the

reference sequence with annotations (colored arrows). Trimmed reads were merged and mapped on to the reference sequence to identify the zero-coverage regions and the overall distribution of coverage (see table 2.23).

- Detecting variation in the mapped data: The possible differences between the sequenced reads in each pool and the reference template along with their position relative to the reference template were determined by alignment to mtLSU rRNA reference sequence (see **attached excel spreadsheet**: merged_SNP_list_from_pool_1_to_pool_3.xls).
  Comparison between assembled reads and the NCBI mtLSU rRNA sequence (AJ973190) identified structural differences that could be structural varioations, misassemblies or sequencing errors. Transversion (A/C, A/T, C/G and G/T) mutations were the most abundant compared to transitions (A/G and C/T).

- Consensus sequences were independently generated for each pool and used for database similarity searches (see attached files: consensus_pool_1.fasta, consensus_pool_2.fasta, consensus_pool_3.fasta and consensus_sequence_merged_pool1-3.fasta). We additionally generated a 2, 607 bps consensus sequence from the three pools (1-3) and used it to search the GenBank for homologs (see section 2.4).

  The consensus sequences of the reads were compared to the GenBank (BLAST) and to the reference sequence in order to build a picture of the structure of the mtLSU rRNA gene, as well as ascertain the species distribution of the closest homologs and improve annotation. The alignment of the reads to the reference sequence is included in the datasets folder:
  i) alignment_file_of_the_reads_to_refseq.pdf
  ii) partial_5_prime_alignment_file_of_the_reads_to_refseq.pdf.
  The read colors are green (forward) and red (reverse) by default. The most plausible explanation for the gaps in the alignment files can be attributed to the high coverage. If you have high coverage in your mapping, you will often find a lot of gaps in the consensus sequence.

1.1 Methods

1.2 Data source

Data was generated in three pools with different tags (MID) and each pool was preprocessed separately followed by merging of the results:

Cecile_pool_1 consisting of 12 sequences
Cecile_pool_2 consisting of 12 sequences
Cecile_pool_3 consisting of 10 sequences

1.3 Extraction of FASTA & Quality files from sff files

An automated sff_extract perl script was used to convert sff files in each pool to fastq and fasta + quality files.

1.4 Quality Control and assessment

Biases in NGS data occurs due to inconsistencies in the quality of reads such as length, quality scores and base distribution. Hence, the raw 454 sequence reads were quality checked using the FastQC program in order to assess the quality of the data and to filter low quality reads

(Barbraham-Bioinformatics, 2009). FastQC is a java program that aims to provide simple ways of doing quality control checks to validate the raw sequenced data and ensures the raw data carries no biases which may affect its usefulness. As a pre-requisite, the adapter sequences and MID tags were trimmed from the raw data using a custom PERL script. The procedure was validated using CLC Genomics Workbench, a commercial software suite for processing NGS data.

Trim settings

• Removal of low quality sequence. (limit = 0.05).
• Removal of ambigious nucleotides: maximal 2 nucleotides allowed.
• Removal of terminal nucleotides: 1 nucleotide from the 5' end and 1 nucleotide from the 3' end.
• Removal of adapter sequences, using the following adapters:
454 Sequence Primer A (CGTATCGCCTCCCTCGCGCCATCAG), strand = Plus, action = Remove adapter, score = [3, 2, 15,2]
454 Sequence Primer B (CTATGCGCCTTGCCAGCCCGCTCAG), strand = Minus, action = Remove adapter, score = [3, 2, 15,2]

1.5 Read alignment and mapping

i) Reference Mapping assembly

The reads were aligned to the reference sequence mtLSU(AJ973190; length 2,607bps) using CLC Genomics Workbench and inGAP pipeline. The mtLSU rRNA reference sequence has three exons separated by two introns (1..40, 1097..1459, 1861..2607).

ii) De novo assembly of the unmapped reads

Contigs were assembled from unaligned reads using CLC Genomics Workbench followed by mapping back of each read onto the assembled contigs.

1.6 Sequence similarity search for unmapped contigs

The contigs generated from the unaligned reads were further screened against the GenBank non-redundant (NR) using BLASTN search algorithm with a criterion imposed at a cut-off expectation value (e-value) of 1e-5. The aim of a BLAST search was to to further characterise contigs as either unique to the source organism (*Glomus intraradices* ??) or as sequencing error. Contigs that did not have a match to any database sequence were classified as species-specific.

2.0 Results:

2.0.1 Quality Control:

The quality control procedure was undertaken to assess properties of the reads such as length, quality scores and base distribution in order to retain high quality reads for downstream assembly or mapping. Input to these tools was sequence data in FASTQ format. The average read length was between 320-400 bps long following trimming.

2.0.2 Pool_1 quality control outputs:

FASTQC and CLC genomics workbench were used to remove or trim primer/adaptor contaminants as well as trim poor quality bases (Table 2.1-2.6).

Table 2.1: Trim Summary for pool 1:

Name	Number of reads	Avg.length	Number of reads after trim	Percentage trimmed
454Reads.MID1 (single)	7,971	370.5	7,891	99%
454Reads.MID10 (single)	1,309	472.2	1,278	97.63%
454Reads.MID11 (single)	3	306.3	3	100%
454Reads.MID12 (single)	3,016	454.5	2,979	98.77%
454Reads.MID2 (single)	6,433	349.1	6,402	99.52%
454Reads.MID3 (single)	18,126	365.3	17,998	99.29%
454Reads.MID4 (single)	9,313	336.8	9,221	99.01%
454Reads.MID5 (single)	15,043	342.9	14,967	99.49%
454Reads.MID6 (single)	23,970	388.5	23,749	99.08%
454Reads.MID7 (single)	9,398	468.4	9,231	98.22%
454Reads.MID8 (single)	18,064	353.5	17,938	99.3%
454Reads.MID9 (single)	536	485.7	521	97.2%

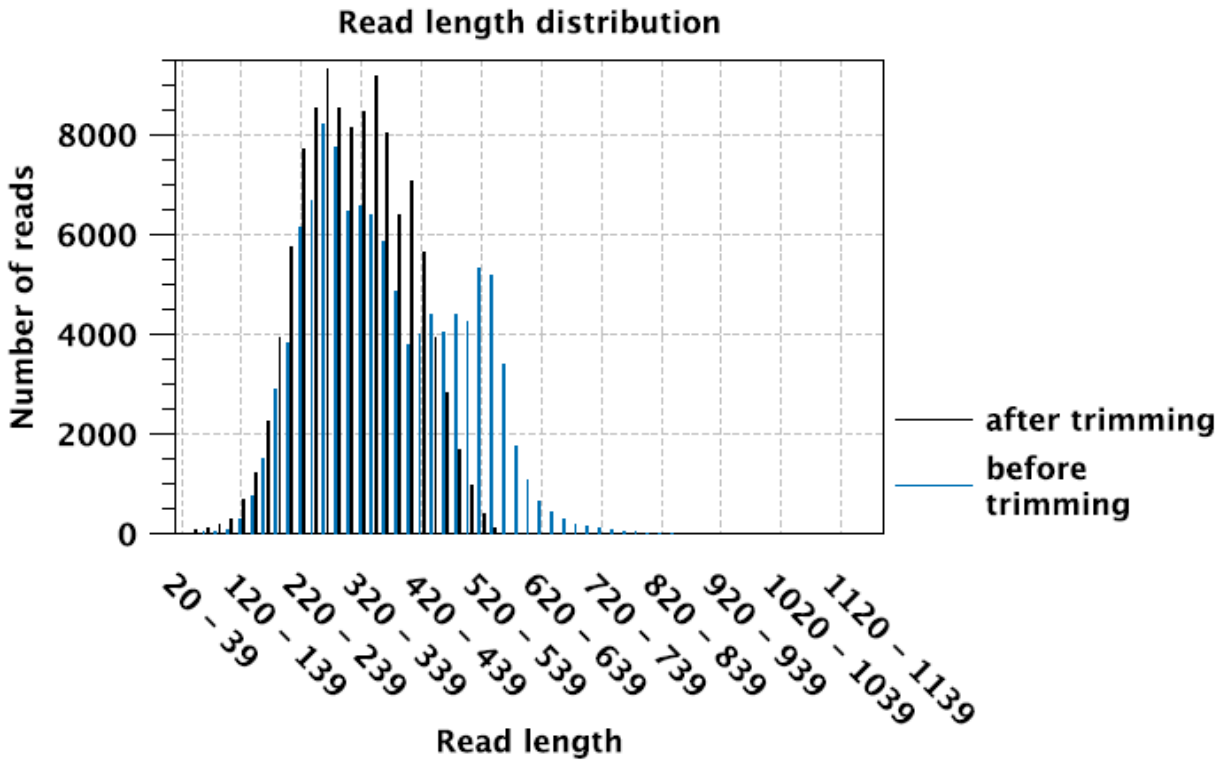

Figure 2.1. Read length before and after trimming

Below is a table with summary statistics for trimming of poor quality reads in pool 1.

Table 2.2: Pool 1 detailed trimming results

Trim	Input reads	No trim	Trimmed	Nothing left or Discarded
Trim on quality	113,182	234	112,948	0
Ambiguity trim	113,182	18,176	94,002	1,004
Trim ends	112,178	0	112,178	0
Adapter trimming	112,178	193	111,985	0

2.0.3 Pool_2 quality control outputs:

Table 2.3: Trim Summary statistics for pool 2:

Name	Number of reads	Avg.length	Number of reads after trim	Percentage trimmed
454Reads.MID1 (single)	8,828	400.5	8,760	99.23%
454Reads.MID10 (single)	762	491.0	751	98.56%
454Reads.MID11 (single)	1,027	534.3	992	96.59%
454Reads.MID12 (single)	792	472.8	782	98.74%
454Reads.MID2 (single)	3,361	391.6	3,343	99.46%
454Reads.MID3 (single)	18,782	419.2	18,578	98.91%
454Reads.MID4 (single)	16,572	442.0	16,298	98.35%
454Reads.MID5 (single)	4,661	424.5	4,607	98.84%
454Reads.MID6 (single)	3,682	381.5	3,655	99.27%
454Reads.MID7 (single)	5,846	458.1	5,755	98.44%
454Reads.MID8 (single)	12,582	421.4	12,437	98.85%
454Reads.MID9 (single)	336	529.4	330	98.21%

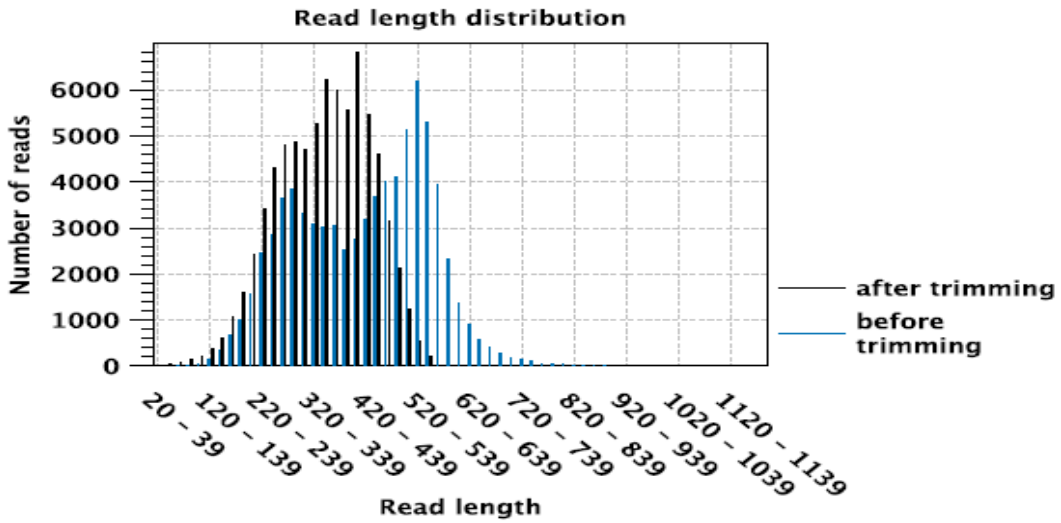

Figure 2.2: Read length before and after trimming

Table 2.4: Pool 2 detailed trim results.

Trim	Input reads	No trim	Trimmed	Nothing left or Discarded
Trim on quality	77,231	213	77,018	0
Ambiguity trim	77,231	15,333	60,955	943
Trim ends	76,288	0	76,288	0
Adapter trimming	76,288	214	76,074	0

2.0.4 Pool_3 quality control outputs:

Table 2.5. Trim Summary for pool 3:

Name	Number of reads	Avg.length	Number of reads after trim	Percentage trimmed
454Reads.MID1 (single)	533	319.7	525	98.5%
454Reads.MID10 (single)	621	480.0	608	97.91%
454Reads.MID2 (single)	1,130	397.8	1,112	98.41%
454Reads.MID3 (single)	29,801	383.4	29,576	99.24%
454Reads.MID4 (single)	22,569	445.3	22,296	98.79%
454Reads.MID5 (single)	3,688	471.0	3,622	98.21%
454Reads.MID6 (single)	3,435	427.7	3,404	99.1%
454Reads.MID7 (single)	335	526.9	326	97.31%
454Reads.MID8 (single)	6,330	438.6	6,024	95.17%
454Reads.MID9 (single)	7	575.0	7	100%

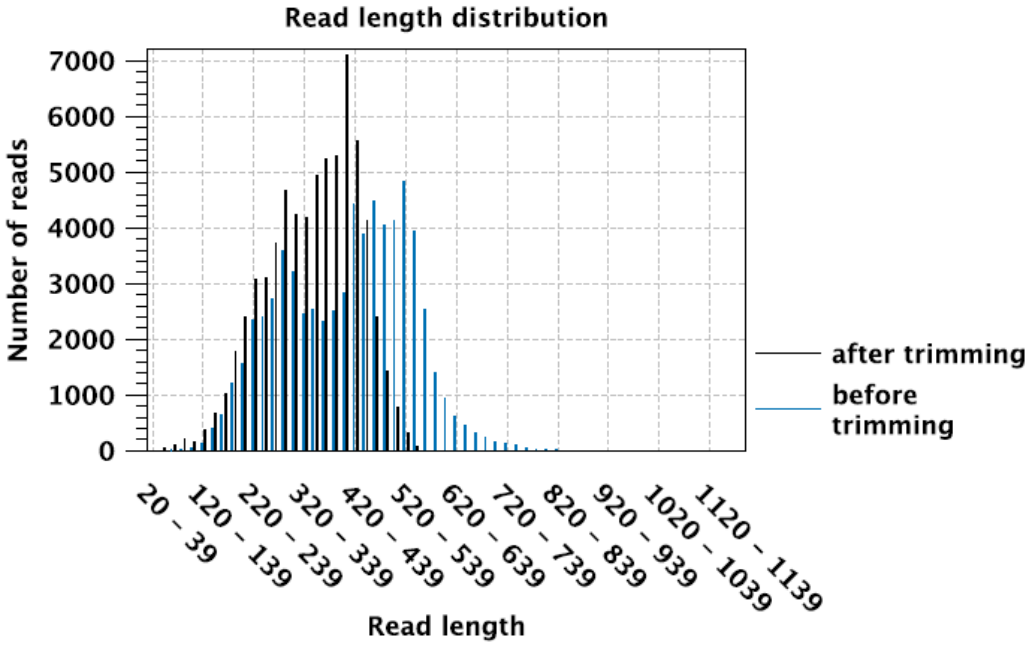

Figure 2.3: Read length before and after trimming

Table 2.6: Detailed trim results.

Trim	Input reads	No trim	Trimmed	Nothing left or Discarded
Trim on quality	68,449	221	68,228	0
Ambiguity trim	68,449	14,124	53,376	949
Trim ends	67,500	0	67,500	0
Adapter trimming	67,500	223	67,277	0

2.1 Read alignment and mapping

2.1.1 Reference mapping assembly

This analysis included calculating the total number of reads aligned and not aligned to the reference mtLSU rRNA sequence.

2.1.2 Cecile_pool_1 mapping:

Table 2.7: Summary statistics for the mapped reads

Read count	78,531
Mean read length	322.53
Total read length	25,328,895

Table 2.8 Summary statistics

Reference count	1
Type	Reference mapping
Total reference length	2,607
GC contents in %	41.85
Total read count	78,531
Mean read length	322.53
Total read length	25,328,895

Table 2.9 mtLSU rRNA coverage

Total reference length	2,607
% GC	41.85
Total consensus length	2,606
Fraction of reference covered	1.00

Table 2.10: Coverage statistics

Total reference length	2,607
Minimum coverage	0
Maximum coverage	15,264
Average coverage	7,276.40
Standard deviation	4,117.69
Minimum excl. zero coverage regions	4
Average excl. zero coverage regions	7,281.99
Standard deviation excl. zero coverage regions	4,114.33

Table 2.11: Coverage regions

Reference Sequence	Feature type	Start position	End position	Length	P-Value
ENA|AJ973190|AJ973190.1	Deletion	2	974	973	0
ENA|AJ973190|AJ973190.1	Amplification	1064	1409	346	0
ENA|AJ973190|AJ973190.1	Amplification	1434	2507	1074	0
ENA|AJ973190|AJ973190.1	Deletion	2510	2607	98	0

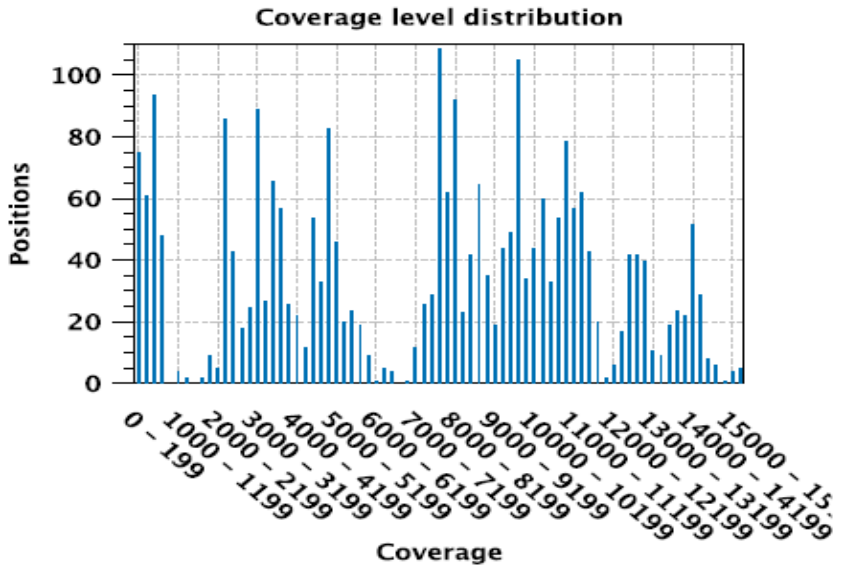

Figure 2.4: Distribution of reads along the reference sequence. Peaks reflect greater coverage, whilst troughs correspond to regions with low coverage (introns).

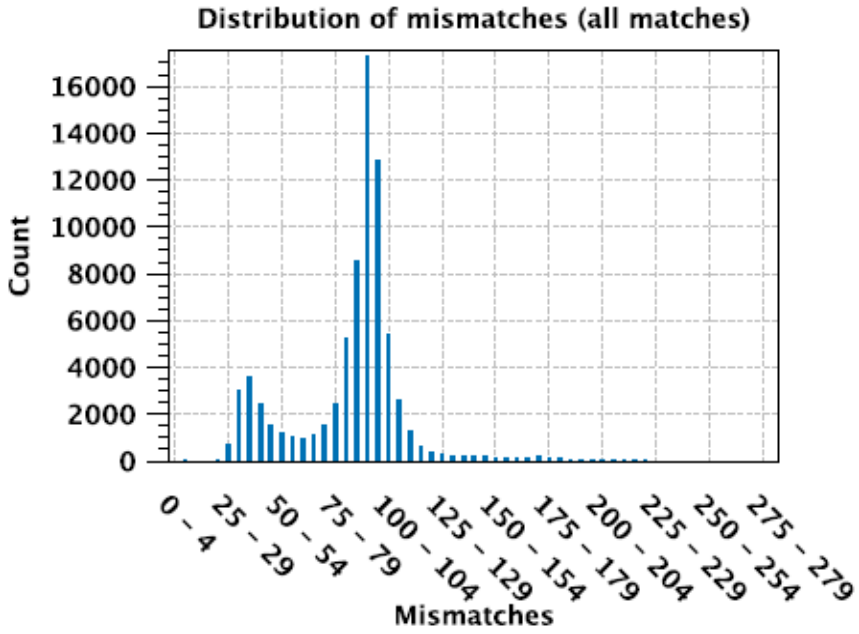

Figure 2.5: Distribution of mismatched bases among reads that mapped to the reference sequence.

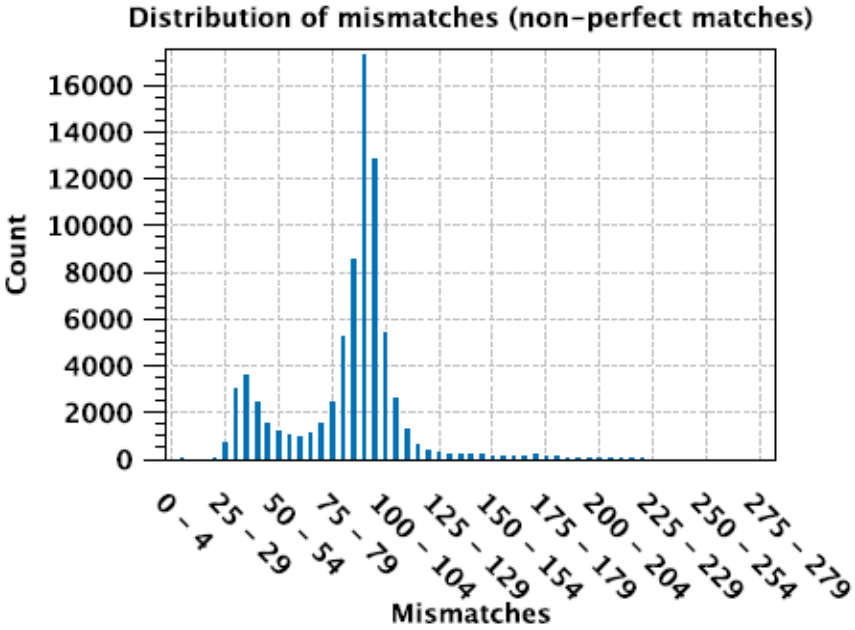

Figure 2.6: Distribution of mismatched bases among unmapped reads.

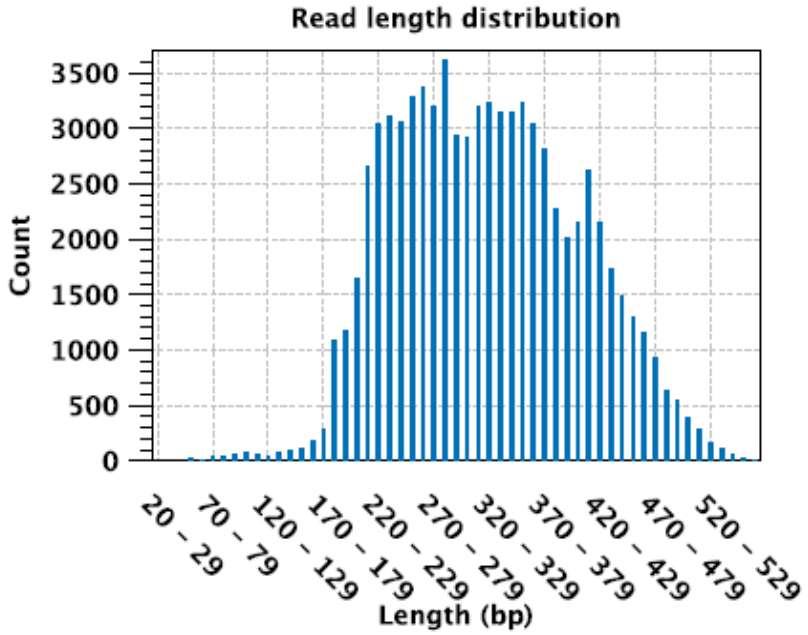

Figure 2.7: Length distribution of the reads.

2.1.3 Cecile_pool_2 mapping:

Table 2.12: Pool 2 summary statistics

Reference count	1
Type	Reference mapping
Total reference length	2,607
GC contents in %	41.85
Total read count	32,163
Mean read length	349.07
Total read length	11,227,120

Table 2.13: Statistics of mapped reads from pool 2.

	Count	Average length	Total bases
Reads	76,288	345.73	26,375,197
Matched	32,163	349.07	11,227,120
Not matched	44,125	343.3	15,148,077
References	1	2,607	2,607

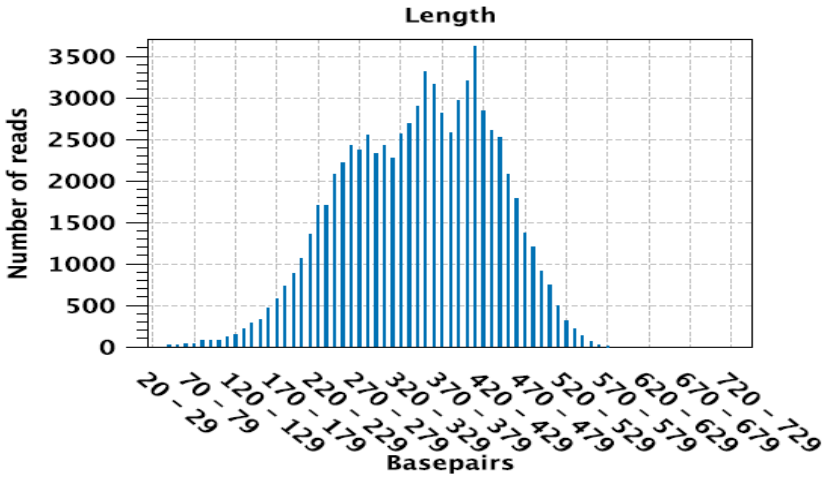

Figure 2.8: Distribution of the length of pool 2 reads that **mapped** to the reference sequence.

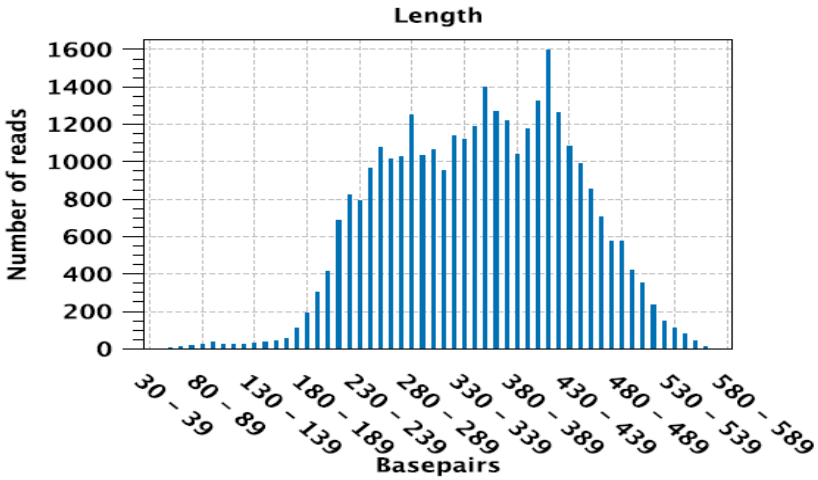

Figure 2.9: Length distribution of the **unmapped** reads from pool 2.

Table 2.14: Percentage of the reference sequence that was covered by pool 2 reads.

Total reference length	2,607
% GC	41.85
Total consensus length	2,610
Fraction of reference covered	1.00

Table 2.15: Coverage statistics

Total reference length	2,607
Minimum coverage	2
Maximum coverage	10,217
Average coverage	3,211.62
Standard deviation	2,180.45

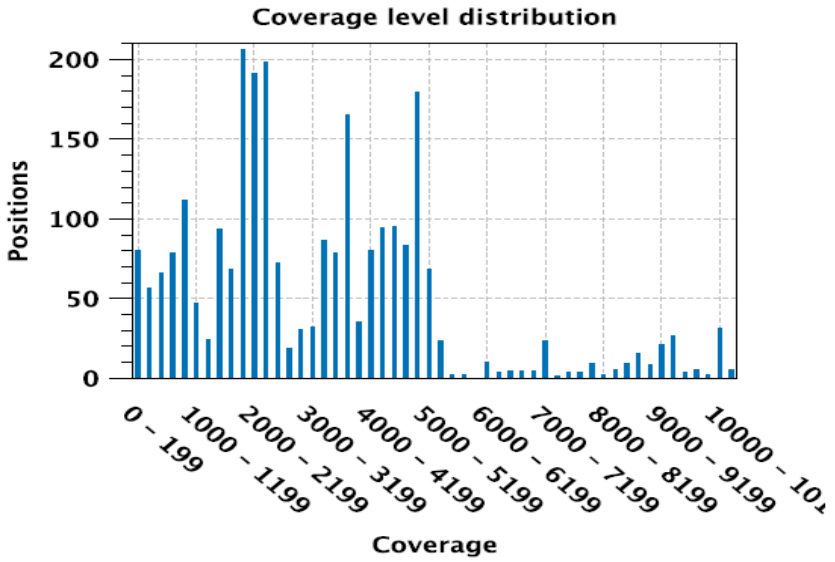

Figure 2.10: Positional coverage of mtLSU mRNA by pool 2 reads.

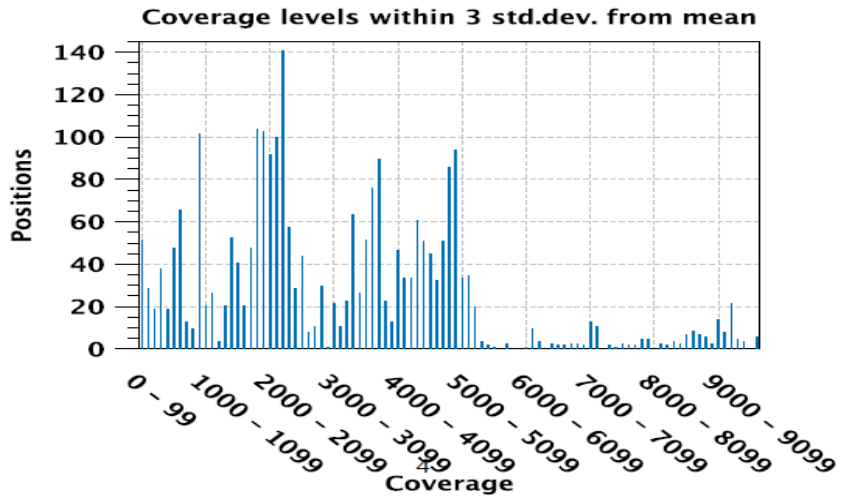

Figure 2.11: Positional mapping of pool 2 reads at 99.73% (3 standard deviation from the mean) coverage.

Table 2.16: Regions of difference/similarity between pool 2 reads and mtLSU rRNA. Regions of deletion correspond to the location of introns within mtLSU rRNA gene.

Sequence name	Feature type	Start position	End position	Length	P-Value
ENA|AJ973190|AJ973190.1	**Deletion**	2	957	956	4.37E-267
ENA|AJ973190|AJ973190.1	Amplification	1017	1544	528	0
ENA|AJ973190|AJ973190.1	**Deletion**	1546	1849	304	4.44E-124
ENA|AJ973190|AJ973190.1	Amplification	1872	2472	601	0
ENA|AJ973190|AJ973190.1	**Deletion**	2509	2607	99	4.97E-187

2.1.4 Cecile_pool_3 mapping:

Table 2.17: Pool 3 summary statistics

Reference count	1
Type	Reference mapping
Total reference length	2,607
GC contents in %	41.85
Total read count	38,886
Mean read length	344.20
Total read length	13,384,577

Table 2.18: Statistics of mapped reads from pool 3.

	Count	Average length	Total bases
Reads	67,500	343.58	23,191,921
Matched	38,886	344.2	13,384,577
Not matched	28,614	342.75	9,807,344
References	1	2,607	2,607

Table 2.19: Percentage of the reference sequence that was covered by pool 2 reads.

Total reference length	2,607
% GC	41.85
Total consensus length	2,588
Fraction of reference covered	0.99

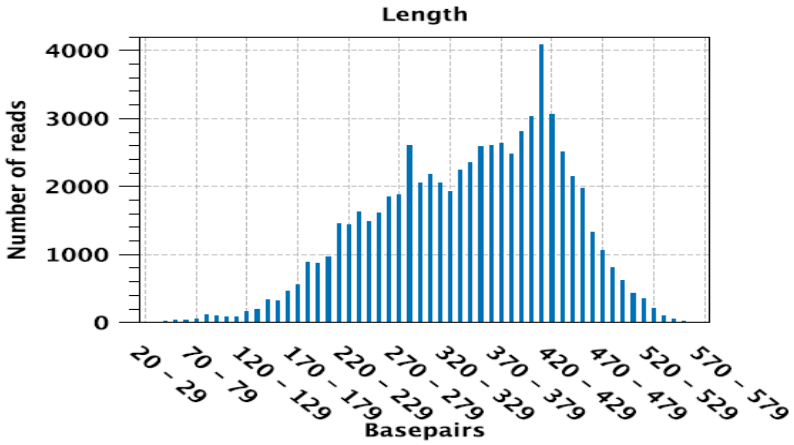

Figure 2.12: Distribution of the length of pool 3 reads that **mapped** to mtLSU rRNA gene.

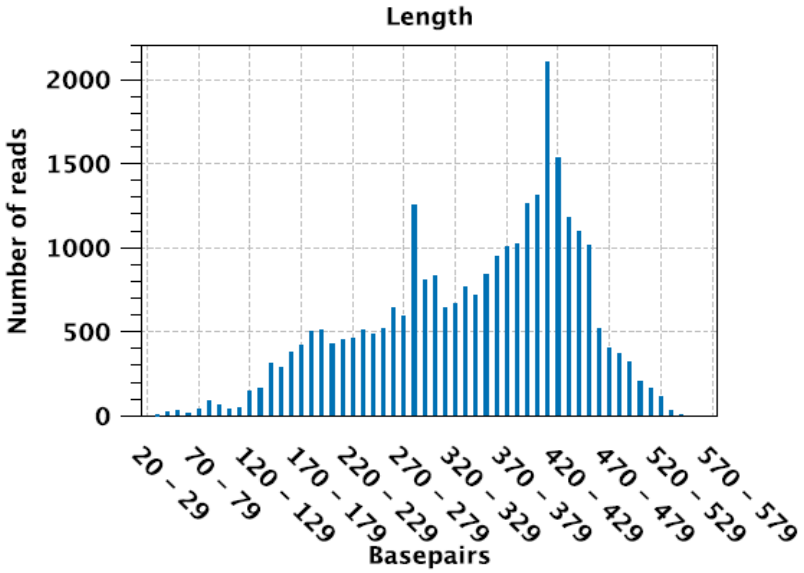

Figure 2.13: Distribution of the length of pool 3 **unmapped** reads.

Table 2.20: Coverage statistics

Total reference length	2,607
Minimum coverage	0
Maximum coverage	10,141
Average coverage	3,925.50
Standard deviation	3,305.58
Minimum excl. zero coverage regions	2
Average excl. zero coverage regions	3,968.12
Standard deviation excl. zero coverage regions	3,297.94

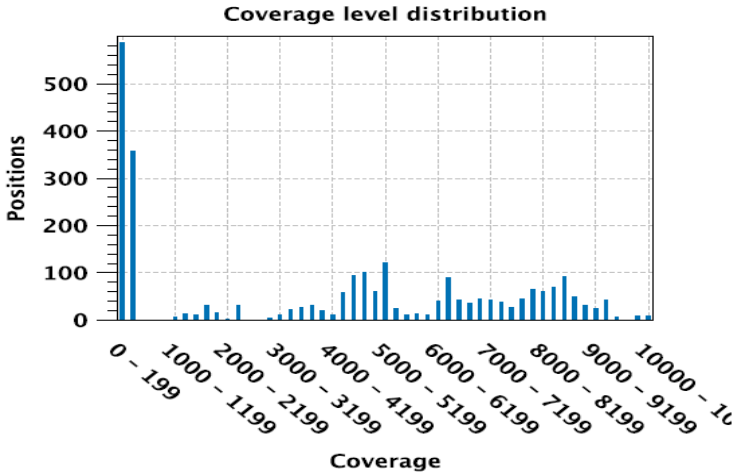

Figure 2.14: Positional coverage of mtLSU mRNA by pool 3 reads.

Table 2.21: Zero coverage regions

Count	1
Minimum length	28
Maximum length	28
Mean length	28.00
Standard deviation	0.00
Total length	28

Table 2.22: Regions of difference/similarity between pool 3 reads and mtLSU rRNA. Regions of deletion correspond to the location of introns within mtLSU rRNA gene.

Sequence name	Feature type	Start position	End position	Length	P-Value
ENA|AJ973190|AJ973190.1	**Deletion**	2	1068	1067	0
ENA|AJ973190|AJ973190.1	Amplification	1081	1404	324	0
ENA|AJ973190|AJ973190.1	Amplification	1431	2508	1078	0
ENA|AJ973190|AJ973190.1	**Deletion**	2511	2607	97	0

Table 2.23. Regions of difference/similarity between merged reads from pool 1-3 and mtLSU rRNA. Regions of deletion correspond to the location of introns within mtLSU rRNA gene.

Sequence name	Feature type	Start position	End position	Length	P-Value
ENA|AJ973190|AJ973190.1	**Deletion**	2	999	998	0
ENA|AJ973190|AJ973190.1	Amplification	1026	1544	519	0
ENA|AJ973190|AJ973190.1	Amplification	1574	2508	935	0
ENA|AJ973190|AJ973190.1	**Deletion**	2510	2607	98	0

2.2 Single nucleotide polymorphism

Structural variation at the nucleotide level such as single nucleotide polymorphisms (SNPs) were called using CLC Genomics Workbench (Table 14).

Table 2.24: Curated genetic variations between pool 3 reads and mtLSU Rrna.

Reference Position	Consensus Position	Variation Type	Reference	Allele	Frequencies	Coverage	Variant #1	Frequency of #1	Count of #1	Type
41	13	SNP	C	T	100	13	T	100	13	Deletion
42	14	SNP	C	T	100	19	T	100	19	Deletion
43	15	SNP	T	A	77.8	18	A	77.778	14	Deletion
108	80	SNP	T	A/T	60.0/40.0	5	A	60	3	Deletion
110	83	SNP	G	A	80	10	A	80	8	Deletion
231	211	SNP	C	T	100	17	T	100	17	Deletion
365	345	SNP	T	A	100	24	A	100	24	Deletion
589	571	SNP	A	T	96	25	T	96	24	Deletion
676	658	SNP	C	T	95.8	24	T	95.833	23	Deletion

2.3 *De novo* assembly of unmapped reads

De novo assembly of 106,312 unmapped reads resulting from alignment to mtLSU rRNA reference sequence generated 248 contigs. The assembly was performed using the following parameter: Minimum contig length 200; word size=20. Reads were mapped back to the contigs with a cut-off imposed at 80% identity.

The N50 of 418 was computed to estimate the quality of the assembly, with the longest contig being 1,219 base pairs (Table 2.25). The contig N50 is a weighted median statistic such that 50% of the entire assembly is contained in contigs equal to or larger than this value.

It (N50) is computed by first ordering all contigs by size and then summing up their lengths until the summed length exceeds 50% of the total length of all contigs. N50 is used widely in genomics, especially in reference to contig or supercontig lengths within a draft assembly to provide a statistical measure of mean length of a set of sequences.

Table 2.25. Contig measurements. The measurements were derived from de novo assembly of unmapped reads.

Length	
N75	355
N50	418
N25	507
Minimum	93
Maximum	1,219
Average	386
Count	248

Table 2.26: Summary statistics of contigs derived from de novo assembly of unmapped reads.

	Count	Average length	Total bases
Reads	106,312	332.47	35,345,270
Matched	46,677	341.76	15,952,338
Not matched	59,635	325.19	19,392,932
Contigs	248	385	95,655

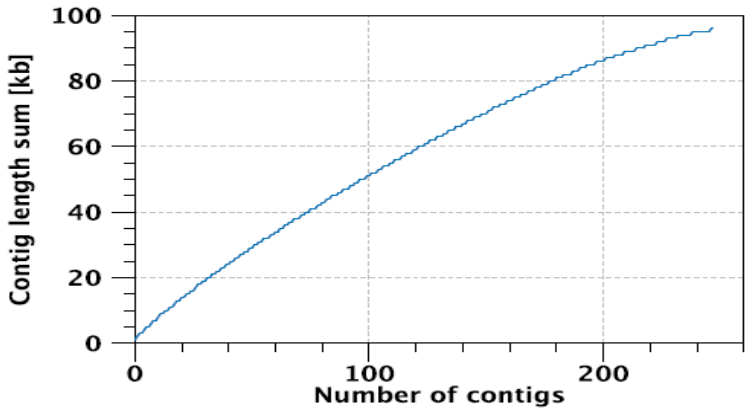

Figure 2.15: Plot of cumulative length distribution of contigs. A distribution plot of the cumulative contig lengths is a visual comparison of assembly results.

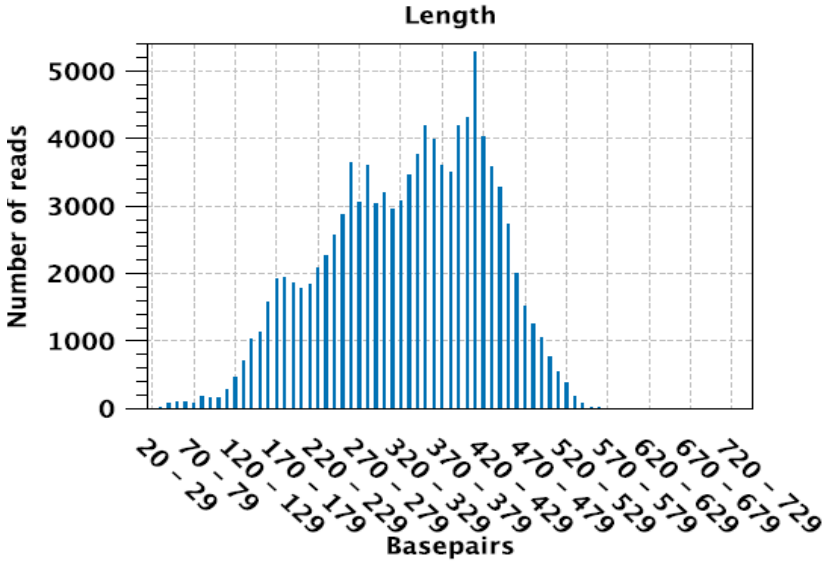

Figure 2.16: Plot of read length distribution of contigs.

2.4 Database similarity searches using unmapped contigs

Database similarity search is a key step to annotate coding sequences in short reads and to identify their function is very challenging because of the size and quantity of the short reads. Traditional alignment programs for the Sanger sequencing are for example BLAST and BLAT but they may not scale well with the NGS reads in terms of the processing time and mapping accuracy. Therefore, consensus sequences were independently generated for each pool and used for database similarity searches (see attached file). We additionally generated a 2, 607 bps consensus sequence from the three pools (1-3) and used it to search the GenBank' non-redundant database for homologs using the BLASTN search algorithm. The cutoff e-value was set at 1e-5 and the number of hits per query sequence was set at 10.
The best matches to the sequenced reads were derived from the *glomeromycetes* taxon with 100% sequence identity (see figure 2.17).

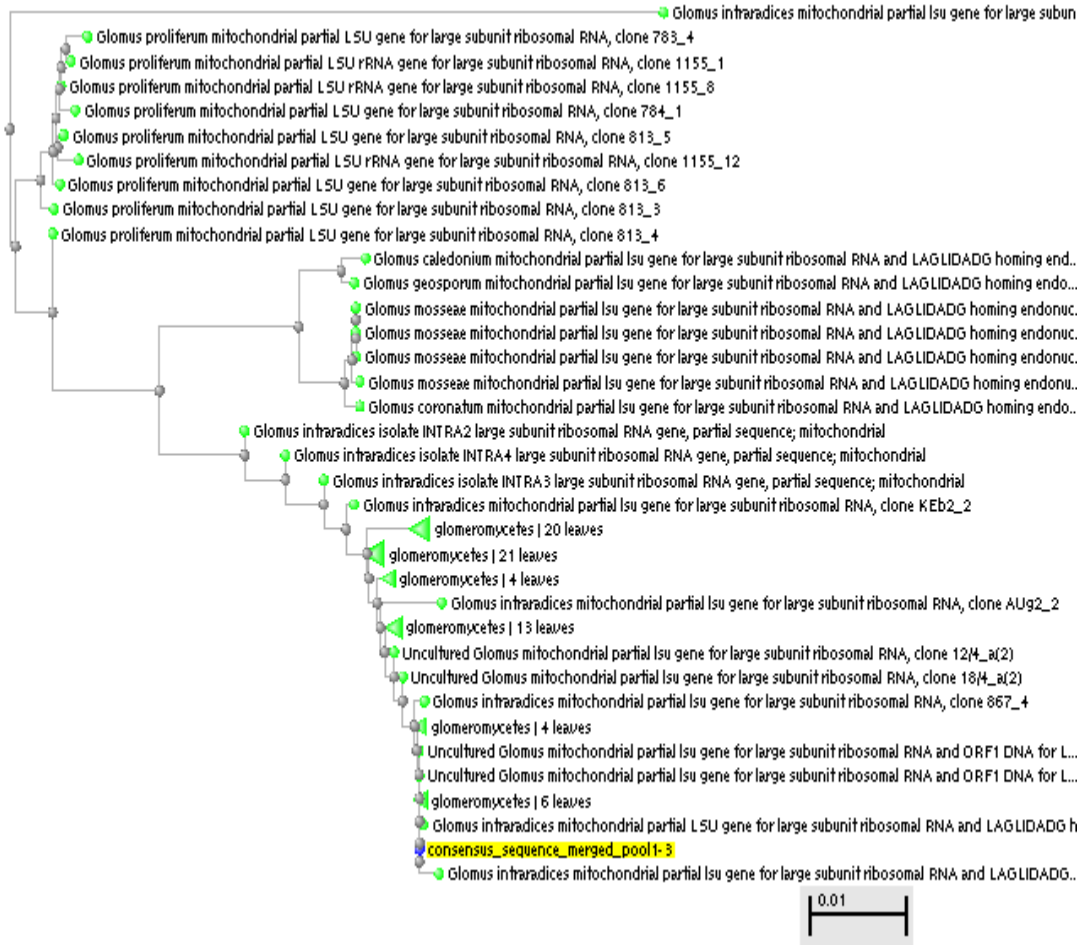

Figure 2.17. Neighbor joining tree depicting taxonomy distribution. The tree was generated following a similarity search at the GenBank using a consensus sequence of the short reads. The query sequence is highlighted in yellow.

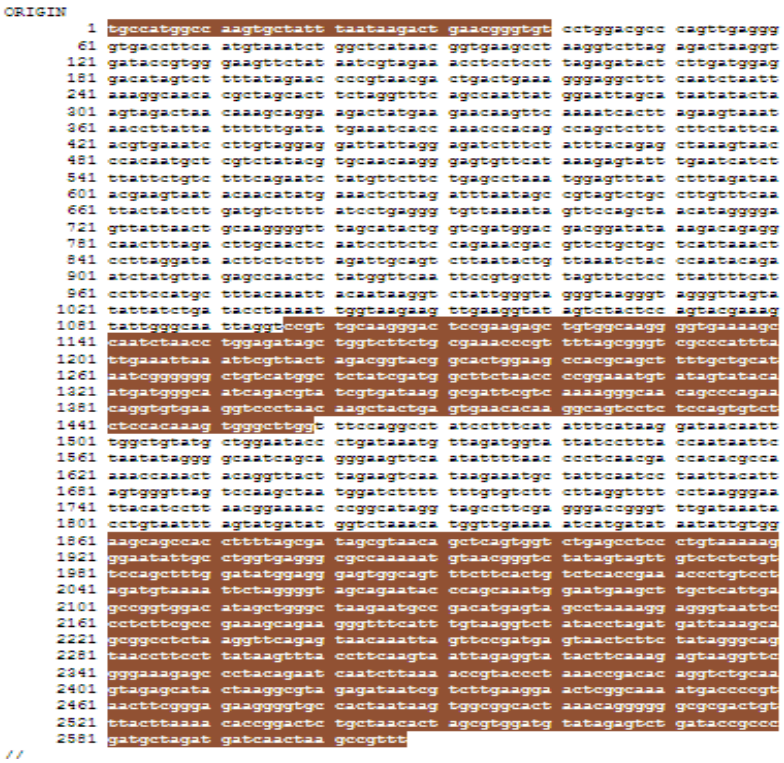

Figure 2.18. Gene structure of mtLSU rRNA (AJ973190: 1..40, 1097..1459, 1861..2607) gene. The gene consists of two introns (**NOT** highlighted) located at position: 41..1096 and 1460..1860.