****

**Spider silk-focused bioinformatic analysis of RNASeq data from *Eratigena atrica***

Name: Mekiran Srividhyadharsan

Project Supervisor: Charlotte Deall

Keywords: *Eratigena atrica*, Major ampullate silk, minor ampullate silk, aciniform silk, tubuliform silk, flagelliform silk, pyriform silk, dragline, spidroin

1. **Introduction**

*Eratigena atrica* (Giant House Spider) belongs to the genus *Eratigena,* in the family *Agelenidae*. Silk is a protein polymer extruded into a fibre and proteins are produced from non-essential amino acids, typically alanine, serine and glycine (Rising et al., 2005). Spider have seven sets of silk-producing glands, each producing a different type of silk (Rising et al., 2005). Silk is an important biomaterial used for commercial applications, due to having specialised properties such as extreme strength and a high capacity to adapt to environmental changes (Rising et al., 2005; X. Hu., et al)

|  |  |
| --- | --- |
| **Silk** | **Core Fibre Proteins** |
| Major ampullate dragline | MaSp1, MaSp2 |
| Minor ampullate | MiSp1, MiSp2 |
| Flagelliform | Flag |
| Aciniform | AcSp1 |
| Tubuliform | TuSp1, ECP-1, ECP-2 |
| Aggregate | Unknown |
| Pyriform | Unknown |

**Table 1:** Spider silk types and their core fibre proteins of *Latrodectus hesperus* (X. Hu et al., 2006)

Four FASTA files of RNA-Sequencing (RNA-Seq) data from *Eratigena atrica* was produced by Charlotte Deall, one file for each sequencing channel. The top line of these reveals the type of sequencing, 1 for single end sequencing and 2 refers to paired end sequencing. R1 is the forward read and R2 is the reverse read. The aim of this research is to carry out de novo transcriptome assembly and carry out analysis on the assembled transcripts using bioinformatic tools.

1. **Materials and Methods**
   1. **Galaxy**

Galaxy is an open source platform used for bioinformatic research and the RNA-Seq data was uploaded to the public server usegalaxy.org (Afgan et al., 2016). Some of their tools were used to complete various tasks that transformed, processed and analysed the data.

**2.1.1 FastQC**

FastQC tool provides quality control checks for the raw sequence data with the information being split into different sections. Each FastQC file was quality checked.

**2.1.2 Trimmomatic**

Trimmomatic is a trimming tool used on the concatenated files and the sequence of trimming operations performed on each files are as follows:

1. Initial Illuminastep: TruSeq3 Adapter sequences, paired-ended, MiSeq and HiSeq
2. LEADING: Cut bases off the start of a read, if below a threshold quality. Set at 20
3. TRAILING: Cut bases off the end of a read, if below a threshold quality. Set at 20
4. SLIDINGWINDOW: Performs a sliding window trimming approach. It starts scanning at the 5‟ end and clips the read once the average quality within the window falls below a threshold. Set at default parameters, 4 bases to average across, 20 is the average quality required.
5. MINLEN: Drop the read if it is below a specified length. Set at 36
6. Output trimlog file
   * 1. **Concatenate**

The four RNA seq files were concatenated into one file, which is required for further analysis of the data. The FastQC and Trimmomatic was run again on the concatenated files and their analysis will be reported in the Results section, not each individual FastQC file. Starting with four files, one for each channel and running FastQC and Trimmomatic allows the data to be checked and cleaned of any errors if necessary, before being grouped together.

* + 1. **Trinity**

Trinity is a de novo transcript assembler found on Galaxy. One challenge with this type of data is that reads are typically smaller than the transcripts they were derived from, so the goal of this tool is re-construct full length transcripts from short reads. Three steps are carried out, with Inchworm’ decomposing reads into overlapping kmers and extending possible kmers, Chrysalis building De Brujin graphs for each gene resulting in thousands of clusters, and Butterfly compacting each De Brujin graph and tracking the path occupied by reads supported by mate pairs. Full length transcripts are outputted. Trinity was performed on the trimmed concatenated files and resulted in R1 single, R1R2 paired and R1R2 unpaired files.

* + 1. **AssemblyPostProcessor**

AssemblyPostProcessor is a tool on galaxy that post-processes the de novo transcriptome assembly into putative coding sequences and their amino acid translations. The transcripts.cleaned.pep file was used for in further analysis. Transdecoder was the coding region prediction method used. The number of sequences found is shown below.

|  |  |  |
| --- | --- | --- |
| **FASTA** | **Trinity (sequences)** | **Transdecoder (sequences)** |
| R1 (single) | 81,162 | 22,469 |
| R1, R2 paired | 98,953 | 24,690 |
| R1, R2 unpaired | 29,499 | 10,375 |

**Table 2:** Number of sequences found after running Trinity and Transdecoder

* 1. **BLAST+ and NCBI**

NCBI and Blast+ was used to create personalised databases that took all the publicly available spider silk data and compared them against the transcripts produced. NCBI search parameters included silk, major ampullate, minor ampullate, aciniform, tubuliform, flagelliform, pyriform, ecribellate, dragline, spidroin and N-term, filtered through animals and excluded shotgun, genome, mitochondrion, plasmids. The purpose of this is that the database created would focus on spider silk. Blastn, Blastp, Blastx, tblastn, tblastx was carried out but the analysis in the results section is on R1R2 paired transcripts compared to the blastn and blastp output. The number of NCBI hits is shown below.

**NCBI hits (search score)**

|  |  |
| --- | --- |
| **Nucleotide (BLASTn)** | **Count** |
| R1 | 743 |
| R1, R2 paired | 952 |
| R1, R2 unpaired | 909 |
| **Protein (BLASTp)** |  |
| R1 | 86067 |
| R1, R2 paired | 95815 |
| R1, R2 unpaired | 40671 |

**Table 3:** Number of NCBI hits for blastn and blastp

* 1. **Benchling: Sequence Alignment**

Benchling was used to compare how well the assembled transcripts as nucleotides or amino acids aligned to the sequences found on NCBI. The template sequence is the NCBI sequence found online, while the aligned sequence is the sequences produced from our de novo transcriptome assembly.

* 1. **ProtParam**

ProtParam was used to find the amino acid composition of any sequences inputted. Using benchling allowed for comparisons to be made between known regions such as the N terminus of the ncbi proteions (with the help of UniProt, an online protein databank to retrieve the ID) and the complementary aligned sequences, determined visually by Benchling.

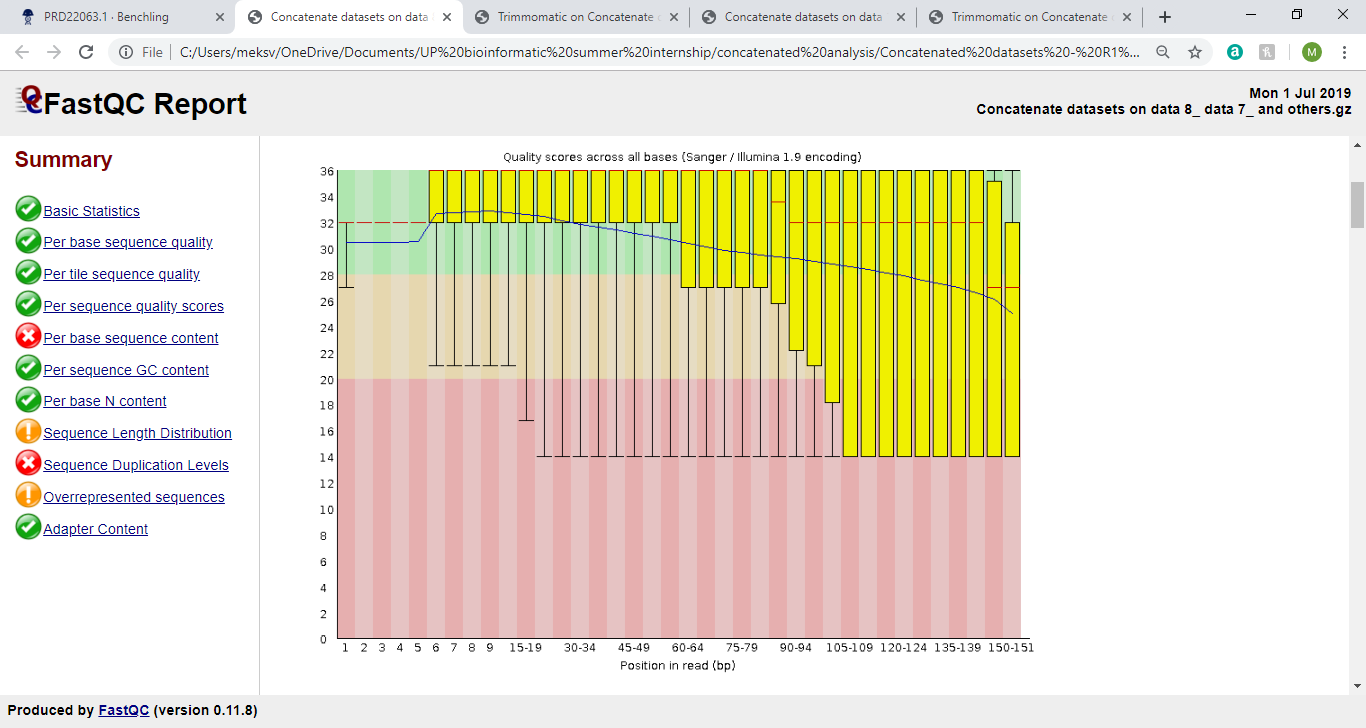
1. **Results**
   1. **FASTQC report: Concatenated and trimmed** 
      1. **Basic statistics**

R1 and R2 should have the same data values given that they are the same reads just read in different directions, R1 is forward and R2 is reverse, and the information shown in Table 1 supports that. The trimming tool, Trimmomatic, has resulted in the same loss of sequence and have slightly altered the sequence length from 35 to 36 (-151).

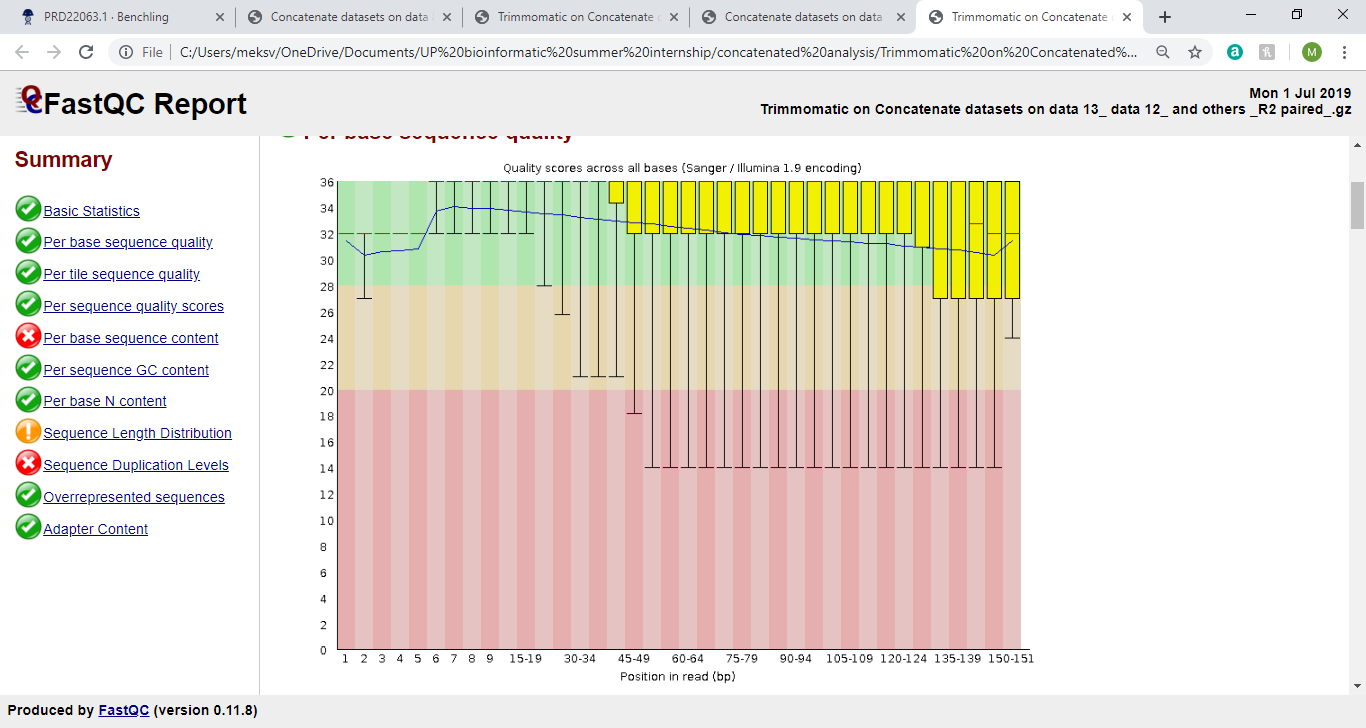
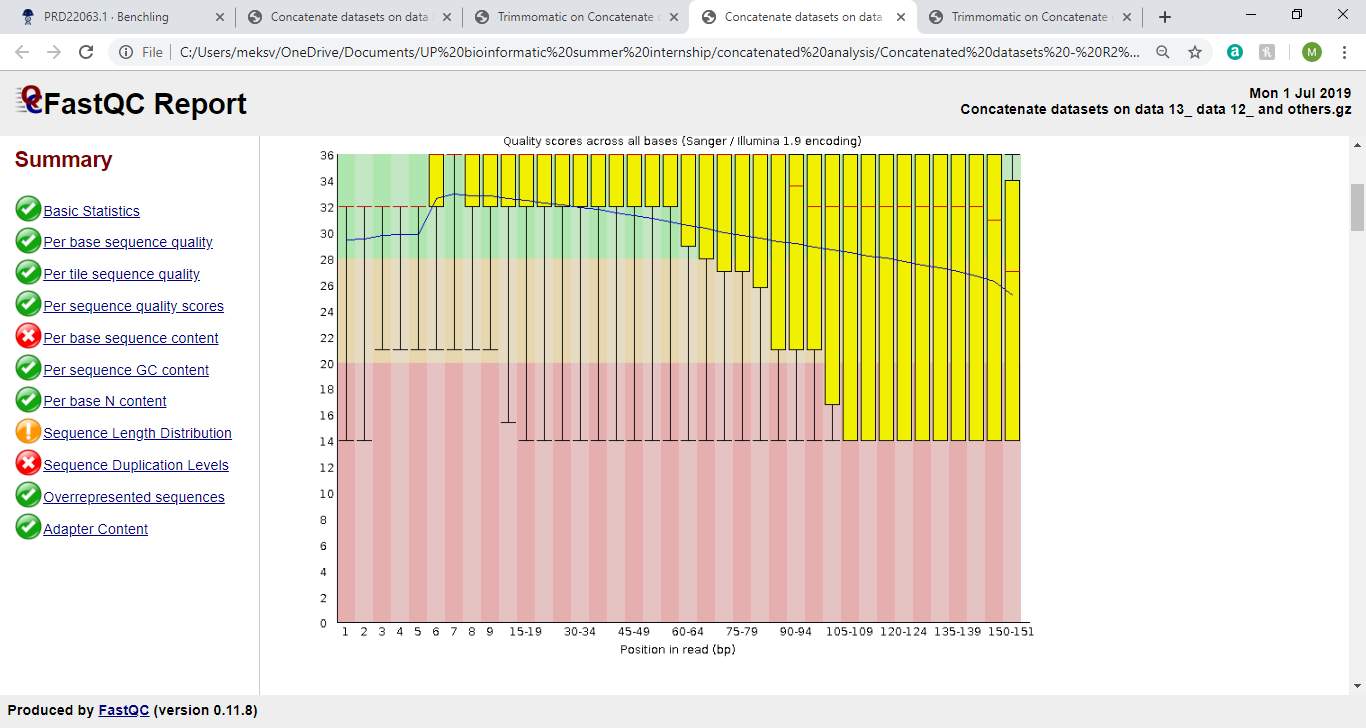
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sequence file** | **Total number of sequences** | **Sequences flagged as poor quality** | **Sequence length** | **%GC** |
| Concatenated R1 | 23956052 | 0 | 35-151 | 39 |
| Trimmed concatenated R1 | 16683006 | 0 | 36-151 | 38 |
| Concatenated R2 | 23956052 | 0 | 35-151 | 39 |
| Trimmed concatenated R2 | 16683006 | 0 | 36-151 | 38 |

**Table 4:** Measures and values from basic statistics sections from FastQC report

* + 1. **Per base sequence quality**



**Figure 1:** Graph from Per base sequence quality section of FastQC report. **Left graph**: Concatenated R1, **Right graph**: Trimmed concatenated R1

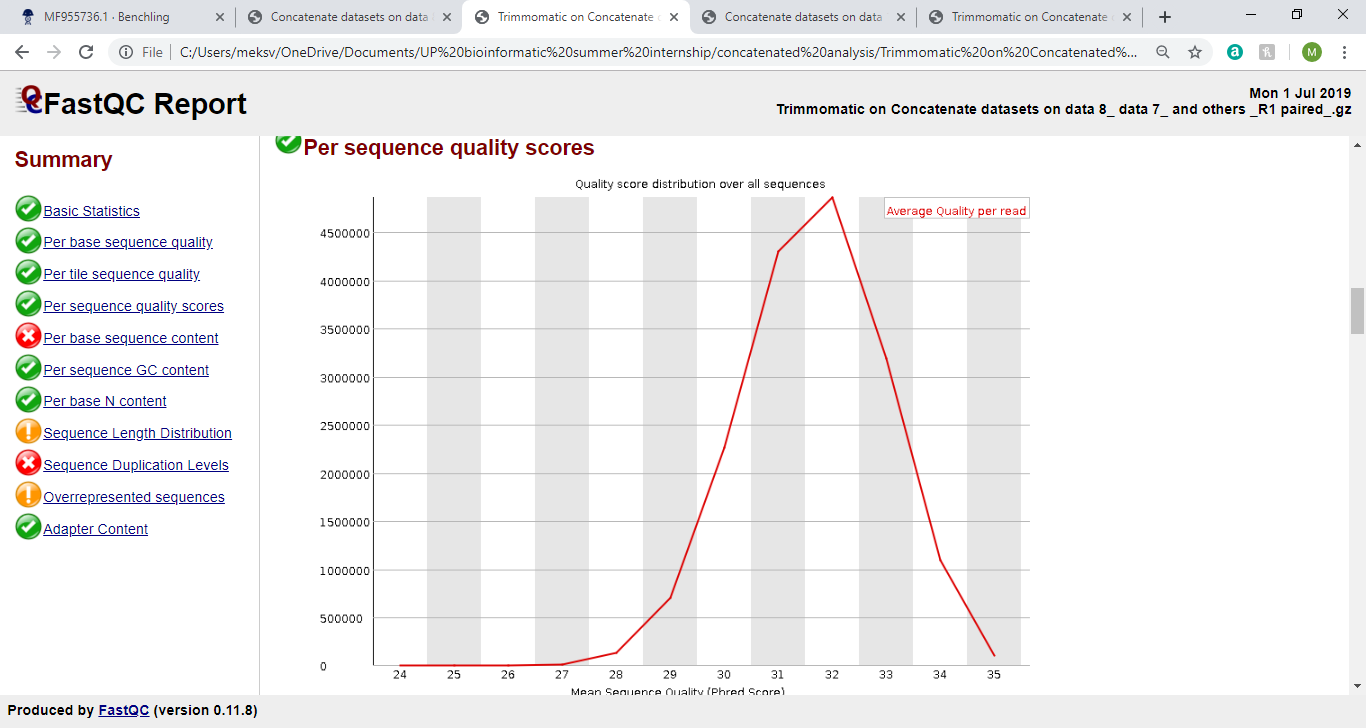
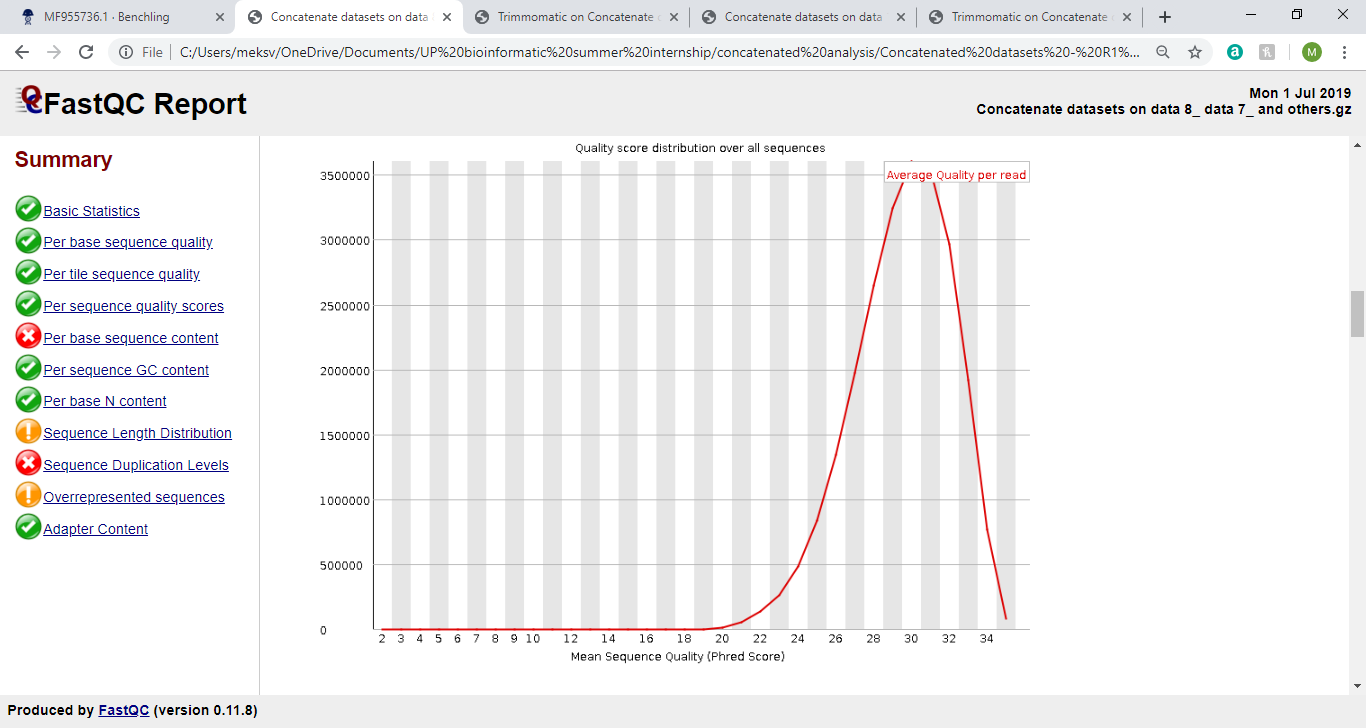


**Figure 2:** Graph from Per base sequence quality section of FastQC report. **Left graph**: Concatenated R2, **Right graph**: Trimmed concatenated R2

Each base position has a BoxWhisper type plot (yellow box), the central red line is the median value, and the blue line represents the mean quality. The y axis on the graph shows the quality scores, so the higher the score the better the base call. From Figures 1 and 2, we can see that trimming has caused the mean quality (blue line) to remain in the green background, meaning that the sequence quality is very good, so sequence quality would negatively affect future analysis. The decline of quality as the bp increases is expected due to the length of our sequences. Unlike other graphs found in different sections of the FastQC report, R1 and R2 have slight differences. The 7-9bp region and the medians at the 150bp regions differ between R1 and R2, most likely due to the direction the sequences are read, but this is not an issue.

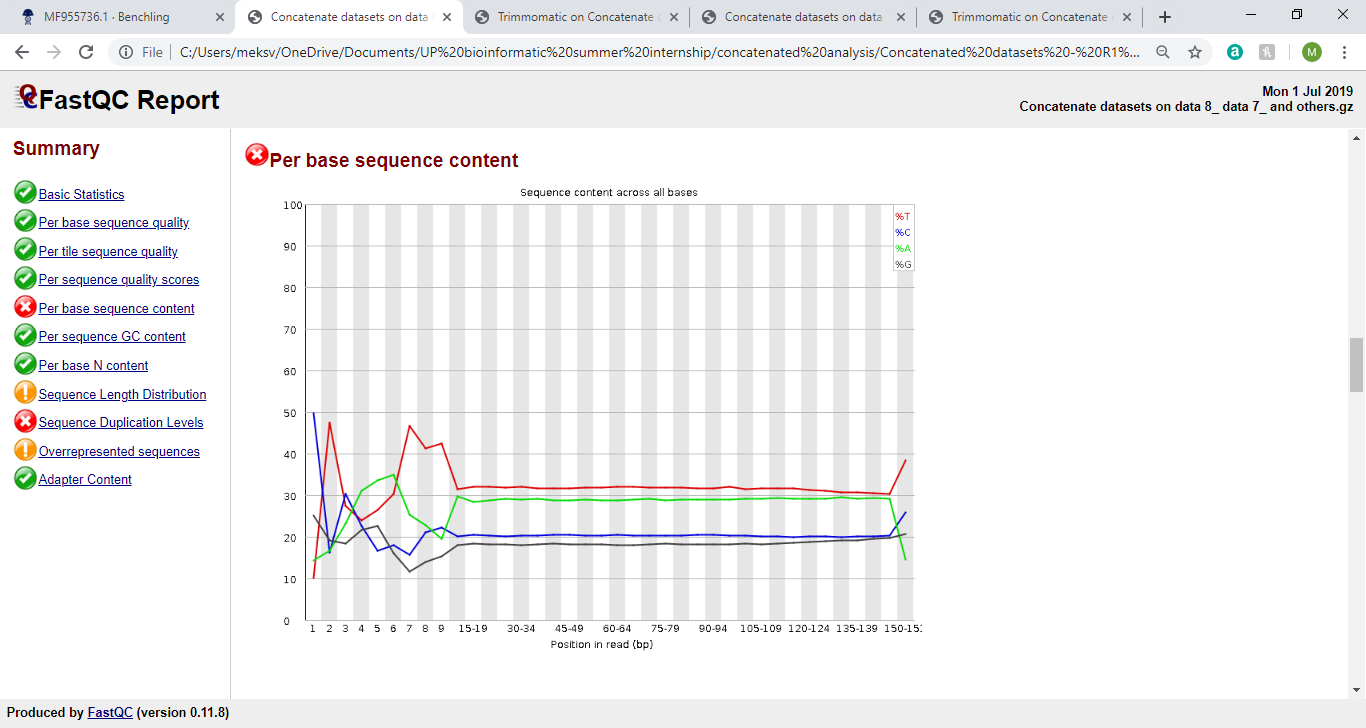
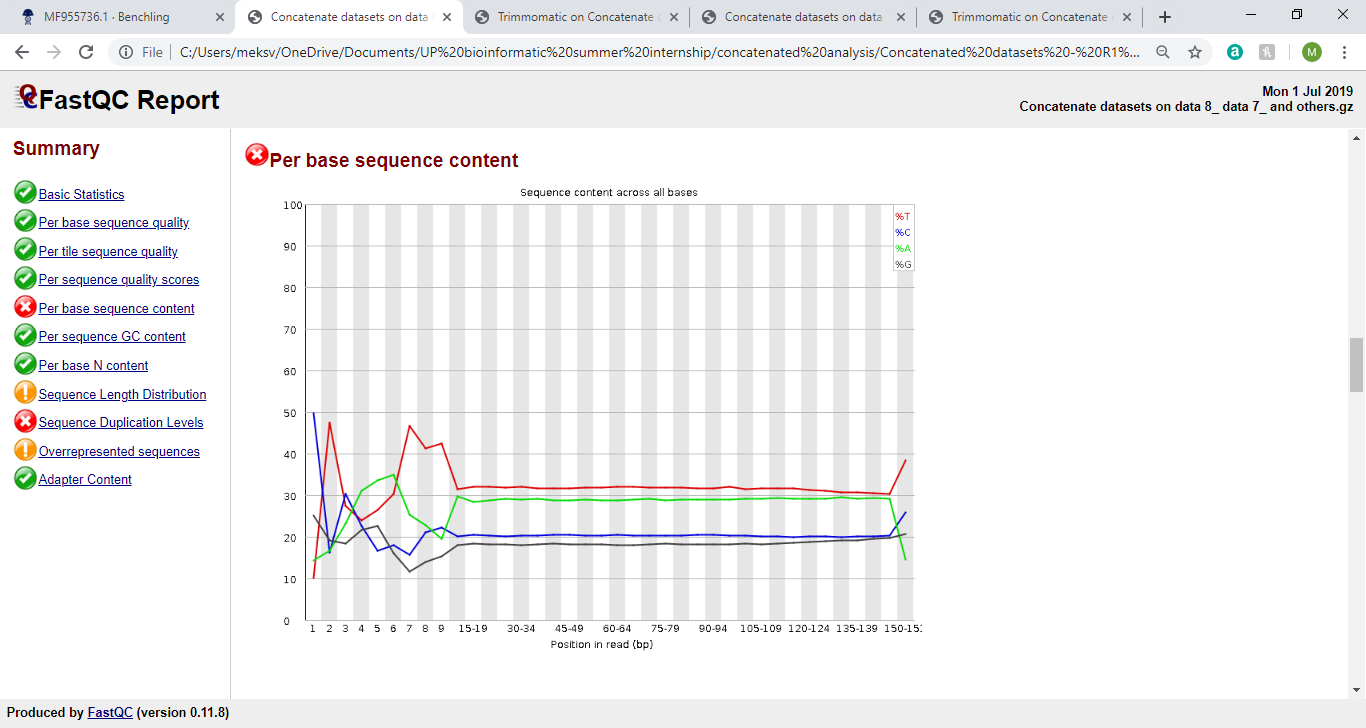
* + 1. **Per tile sequence quality**

This section only appears if you’re using Illumina library which retains its original sequence identifiers. There is a flowcell tile from where each read came. The graph lets you look at quality scores from each tile across all your bases to see if there is a loss in quality associated with only one part of the flowcell.Cold to hot scale, with cold colours being positions where the quality was at or above the average for the base in the run. Hotter colours mean that tile is in a worse quality than other tiles for that base. A good plot should be blue all over.All graphs in the four files were blue all over, and the trimming has resulted in higher quality in tiles, as there were less light blue areas, although the change is quite slight. (Graph not shown here due to difficulty to visualise changes, unless size is increased).

**3.1.4 Per sequence quality score****s**

**Figure 3:** Graph from Per sequence quality scores section of FastQC report. **Left graph**: Concatenated R1, **Right graph**: Trimmed concatenated R1

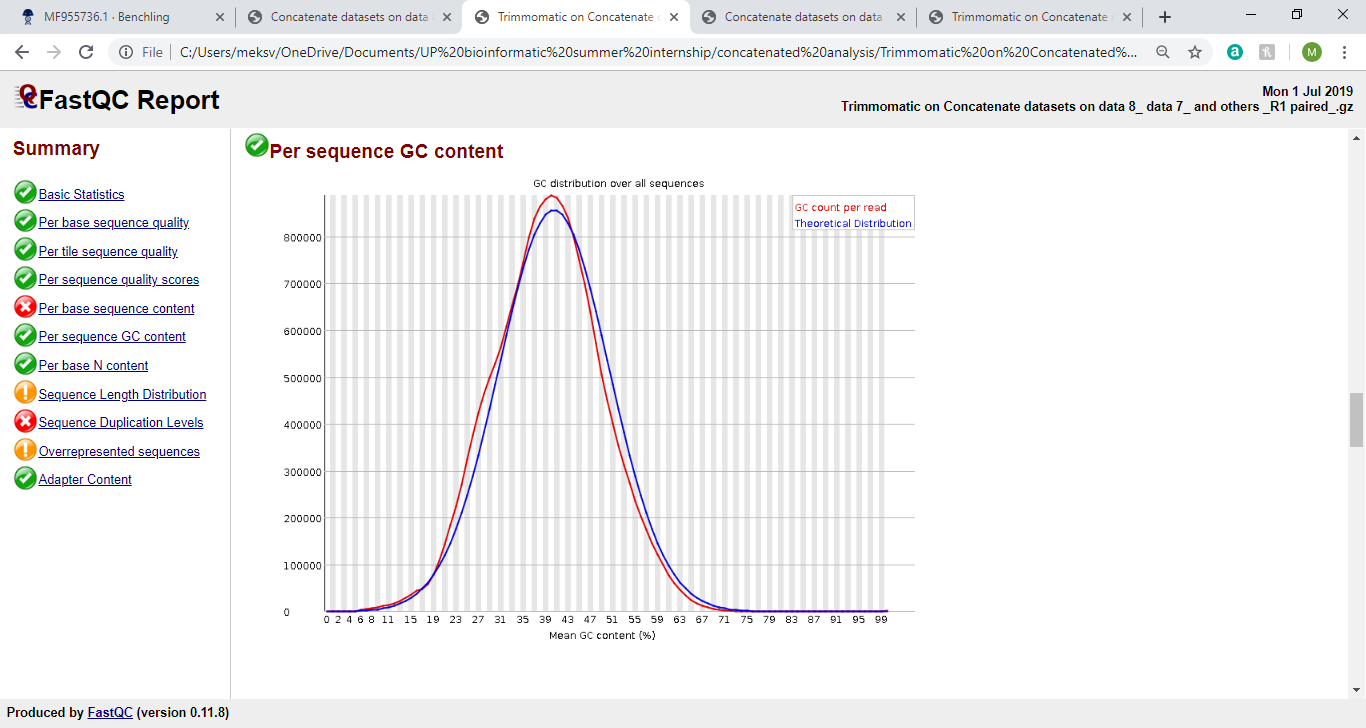
From Figures 3, the peak passes the 27 phred score for all graphs, which means that the sequence quality is good. We can see that for R1, the average quality per read peaks is around 30 phred score to 3500000 sequences, but after trimming, the average quality peaks at 32 phred score to more than 4500000 sequences, so trimming has improved the quality. R2 graphs are the same

**3.1.5 Per base sequence content**

**Figure 4:** Graph from Per base sequence content section of FastQC report. **Left graph**: Concatenated R1, Right graph: Trimmed concatenated R1

From Figure 4, the fluctuations found in the first 9bp of the graph is expected as the data is from an RNA-Seq libraries and is caused by hexamers. Trimming was not expected to fix the issue and it clearly has not, but since the lines eventually plateau, there should be no issue in our downstream analysis. Based on the lines on the graph, there are more adenine (A) and thymine (T) bases (around 30% each), than cytosine (C) and guanine (G) bases in our sequences (around 20% each)

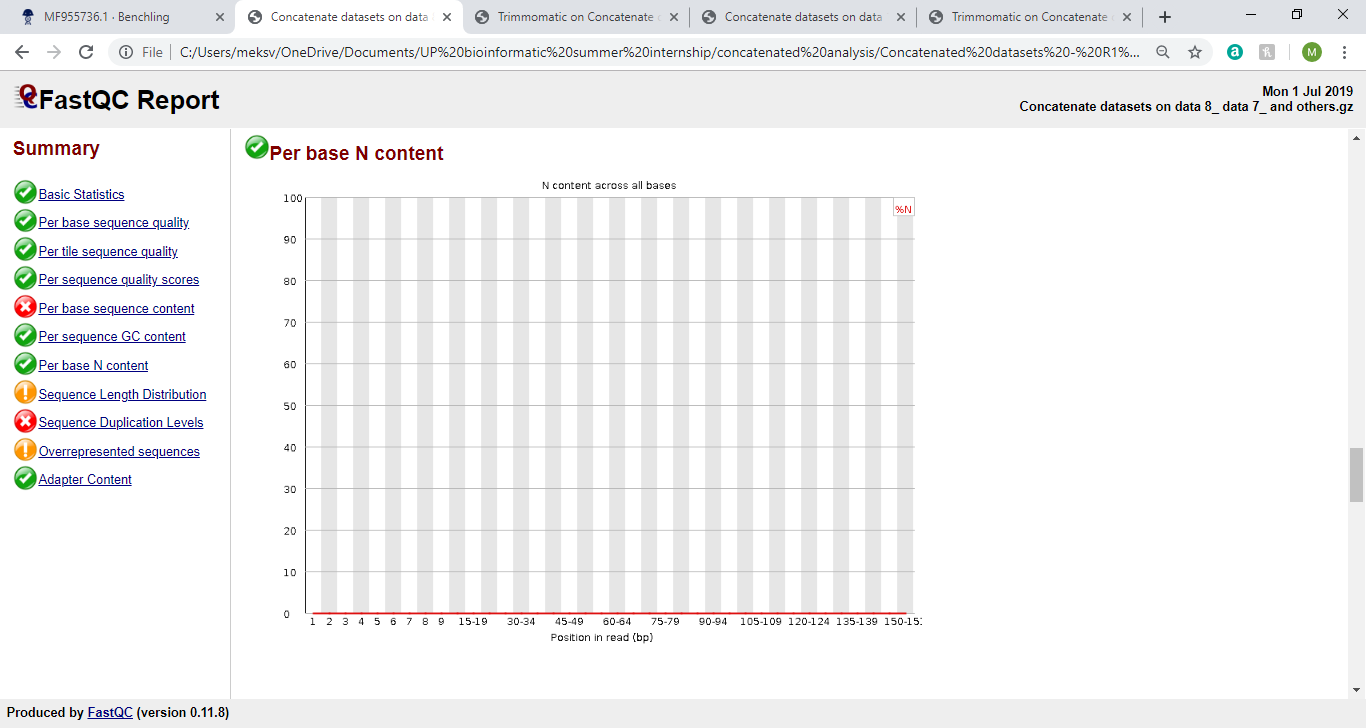
**3.1.6 Per sequence GC content**



**Figure 5:** Graph from Per base sequence content section of FastQC report for Concatenated R1

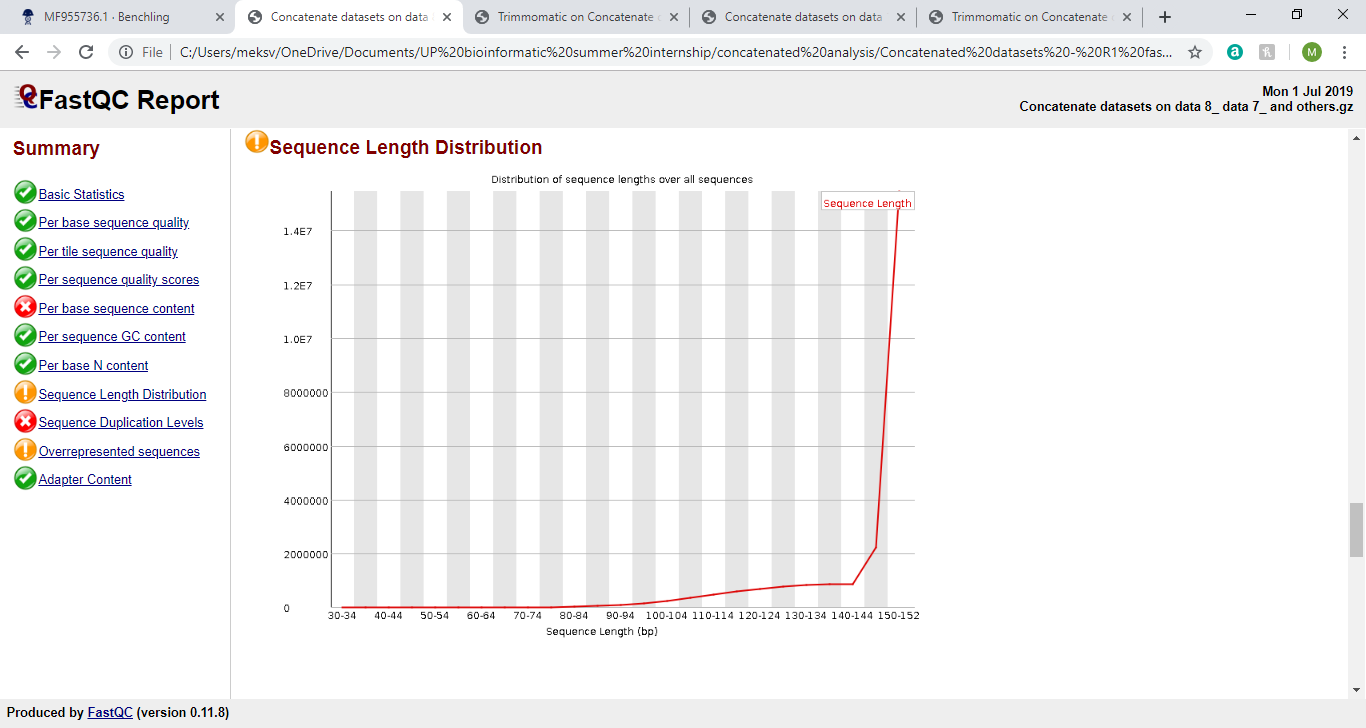
The theoretical distribution (blue line) matches really well to the GC count per read (red line), which is good. The 40% mean GC content matches with previous data. Graphs from other FastQC reports are identical or have very little difference.

**3.1.7 Per sequence N content**



**Figure 6:** Graph from Per sequence N content section of FastQC report for Concatenated R1

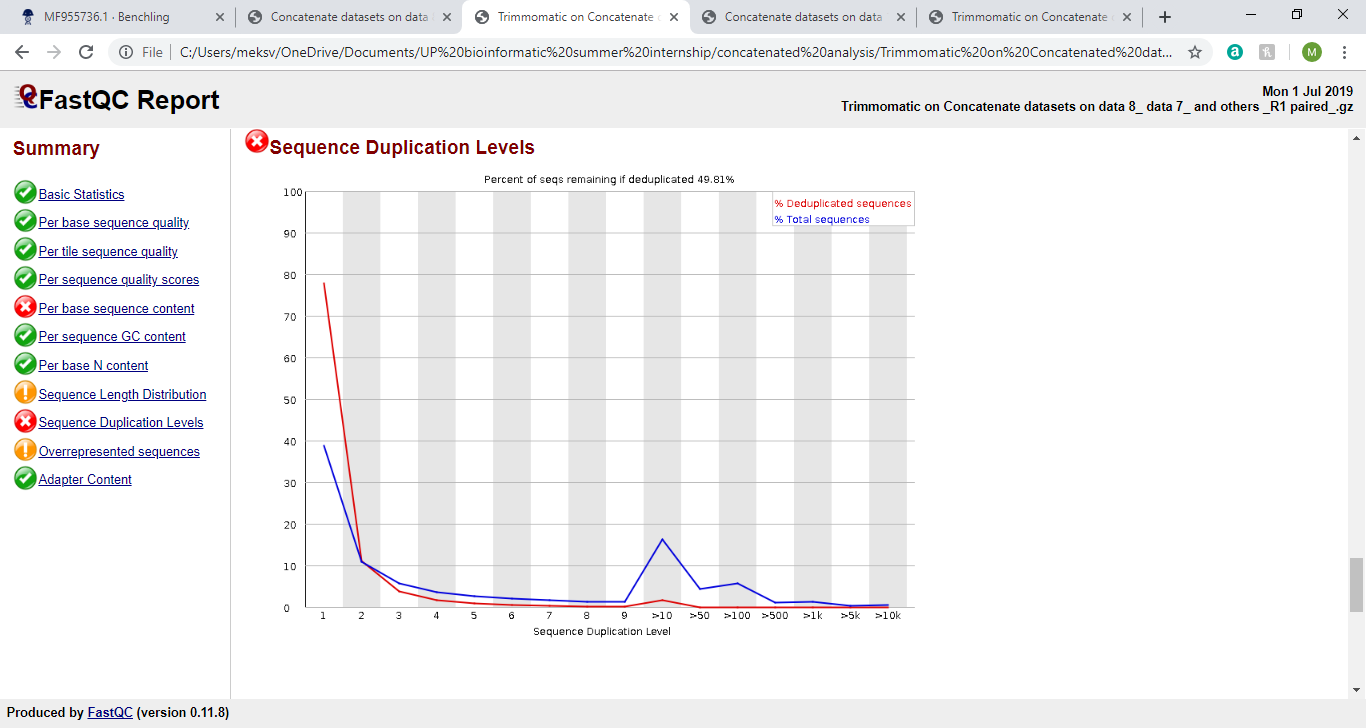
A very low proportion of Ns is good as it means that the sequencer has identified a base at most positions on the sequence and there should only be a few Ns. This graph matches other found in other FastQC reports.

**3.1.8 Sequence Length Distribution**

**Figure 7:** Graph from Sequence Length Distribution section of FastQC report. **Left graph**: Concatenated R1, **Right graph**: Trimmed concatenated R1

All of our sequences will not have the same length, so the warning for this section in FastQC report is due to that. The change in sequence length distribution after trimming is expected and good as it is evidence that the trimming has worked.

**3.1.9 Duplicate sequences**



**Figure 8:** Graph from Duplicate sequences section of FastQC report. **Left graph**: Concatenated R1, **Right graph**: Trimmed concatenated R1

The sequence duplication level peaks at 10, and while this raised an error in all FastQC level, this is not an issue as RNA sequences will have a few duplications as more RNA transcripts means that more proteins will be produced.

**3.1.10 Overrepresented sequences**

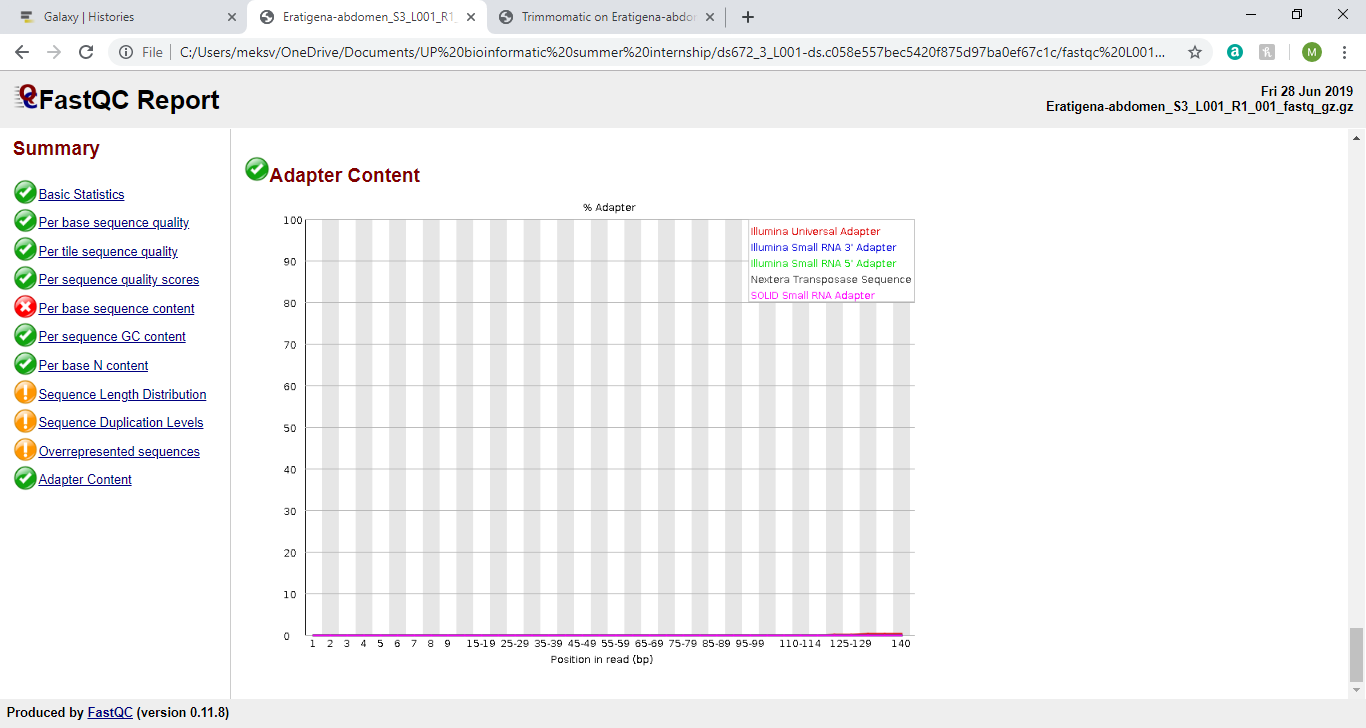
R2 had no overrepresented sequences. R1 had three shown in table below and trimming did not remove them. The first two sequences are partial sequences for 16S ribosomal RNA isolated from our species of interest *Eratigena atrica*, with the sequence ID: KY015700.1. The last sequence is a partial sequence for 16S ribosomal RNA and a mitochondrial gene for a mitochondrial product for *Tegenaria saeva*, with a sequence ID of AY138825.1. Since these are all from spiders and mitochondrial products will be cleared in later analysis as it is not the focus of this investigation, this is not an issue.

|  |  |  |  |
| --- | --- | --- | --- |
| Sequence | Count | Percentage | Possible Source |
| CTAGTCTTTAATTAAAAGACAAATGA  TTATGCTACCTTAGCACAAATTAA | 29262 | 0.12214867458127072 | No Hit |
| GTTTAATTGTTAAATCTTTCGTTCTAG  TCTTTAATTAAAAGACAAATGAT | 28787 | 0.1201658770819165 | No Hit |
| CTTTTTCTTTAATAAGATCTTTGAGAAA  TTATTGCGCTGTTATCCCTATG | 24114 | 0.10065932399879579 | No Hit |

**Table 5:** Details on the overrepresented sequences

Det

**3.1.11 Adapter Content**

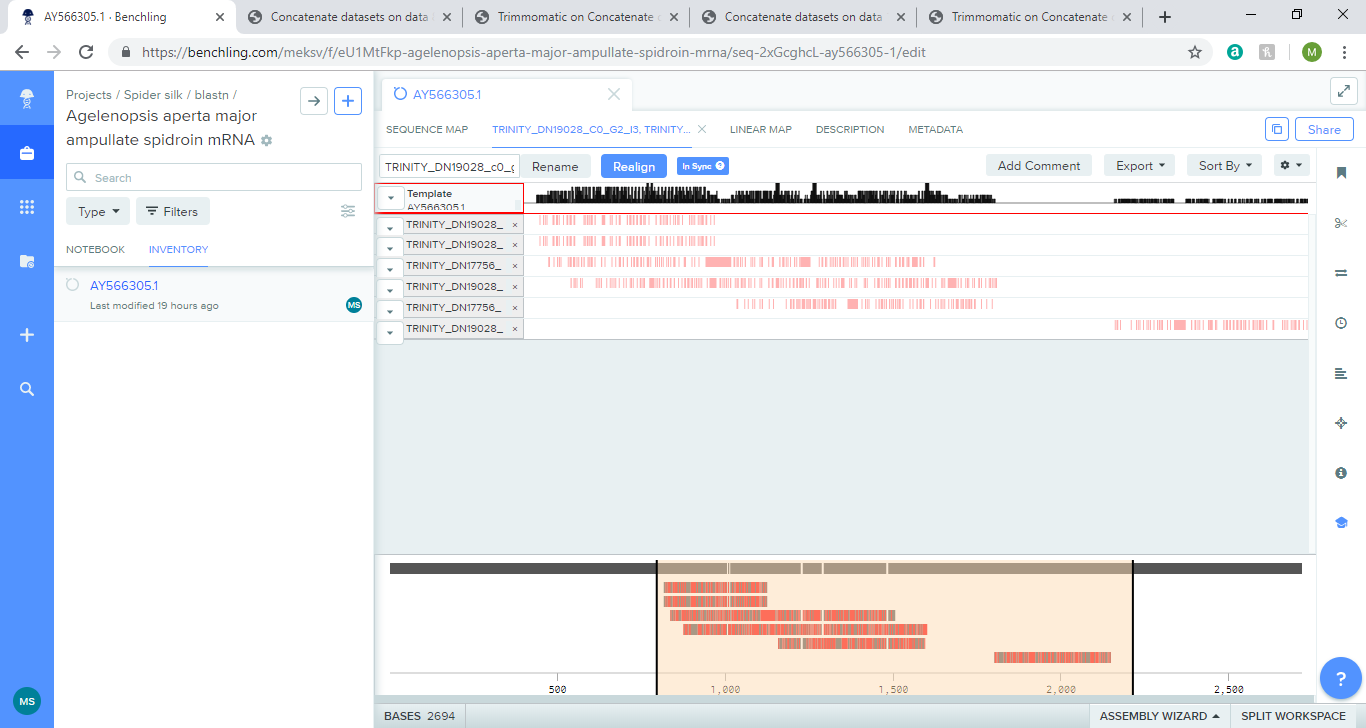


**Figure 9:** Graph from Adapter Content section of FastQC report for Concatenated R1

This is the same in all FastQC files and it shows that only a SOLID Small RNA adapter was used in the RNA sequencing.

* 1. **Blastn**

From the blastn output from BLAST+, three sets of sequences that are directly related to silk have been taken from NCBI and sequence aligned to various query sequences from the assembled transcripts produced by Trinity.

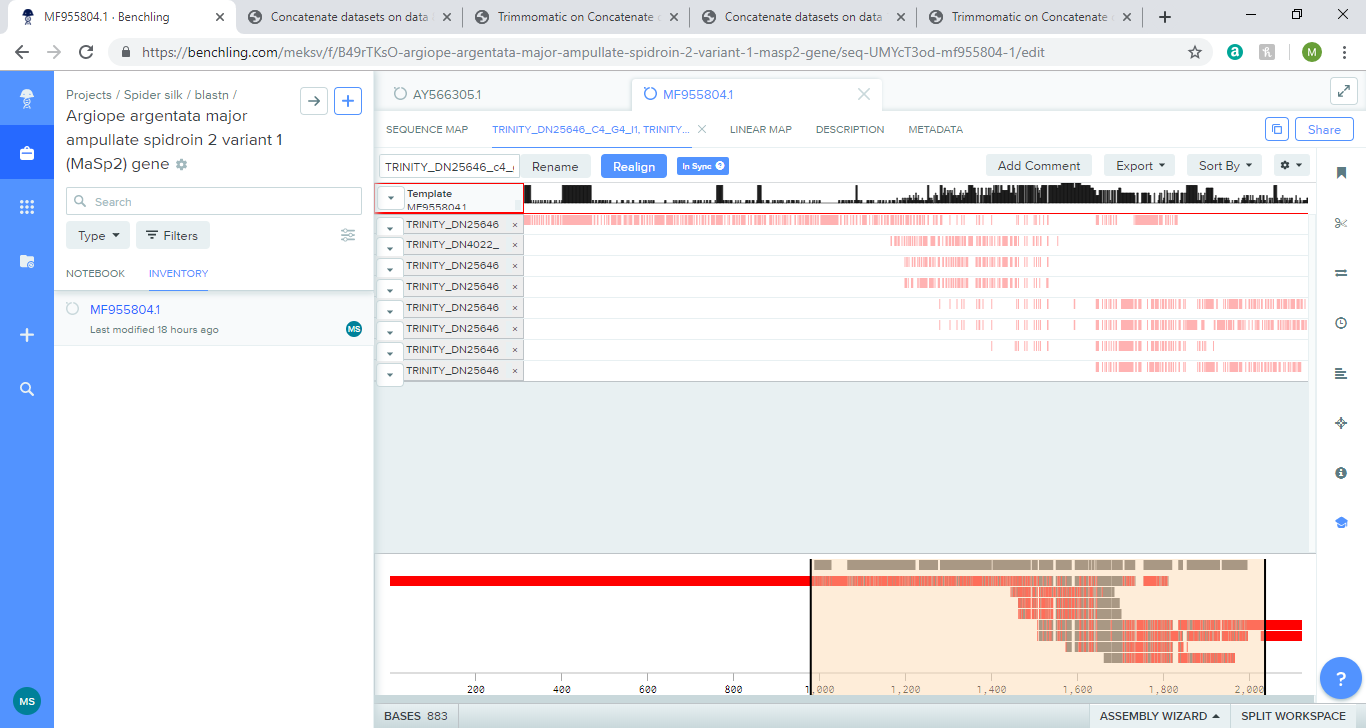
**3.2.1 Agelenopsis aperta major ampullate spidroin mRNA**

**Figure 10:** Sequence alignment for Agelenopsis aperta major ampullate spidroin mRNA. Red lines mean different nucleotide. Spaces between red lines means same nucleotide.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Query ID** | **E value** | **Bit score** | **Percenatage of identical matches** | **Alignment length** | **Number of mismatches** |
| TRINITY\_DN19028\_c0\_g2\_i3 | 2.91E-22 | 100 | 87.356 | 87 | 11 |
| TRINITY\_DN19028\_c0\_g2\_i9 | 2.91E-22 | 100 | 87.356 | 87 | 11 |
| TRINITY\_DN17756\_c0\_g1\_i3 | 2.82E-20 | 95.3 | 86.207 | 87 | 12 |
| TRINITY\_DN19028\_c0\_g2\_i8 | 1.94E-22 | 102 | 71.888 | 466 | 96 |
| TRINITY\_DN17756\_c0\_g1\_i4 | 1.89E-15 | 78.7 | 77.536 | 138 | 25 |
| TRINITY\_DN19028\_c0\_g2\_i6 | 1.15E-21 | 99 | 87.209 | 86 | 11 |

**Table 6:** Blastn for paired sequencing for Agelenopsis aperta major ampullate spidroin mRNA. Some query IDs have multiple rows in blastn document that differ in final 5 column headings only. Out of those multiple rows, the data shown is the row that is ordered first.

A high percentage of identical matches ranging from 71.8% to 87.4% (3sf), means that there is a high level of sequence similarity. Sequences are translated in forward direction (5’ to 3’), due to positive number and start with 1st reading frame based on data. Four of the six sequences begin their alignment with the subject at 786bp and ending at 872bp, with one sequence starting at 787bp instead (and ending at 872bp, as well). TRINITY\_DN19028\_c0\_g2\_i8 has a higher number of mismatches than other sequences, but the alignment length is also bigger so this may not be an issue.

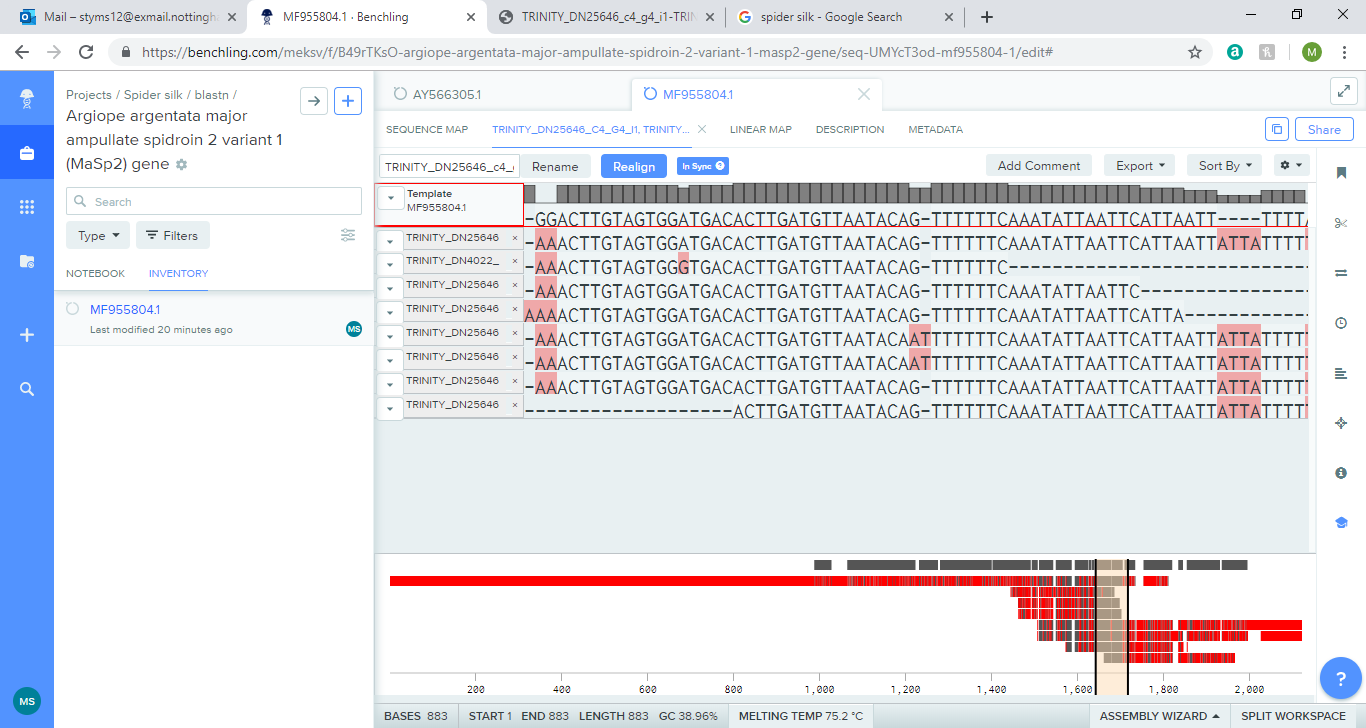
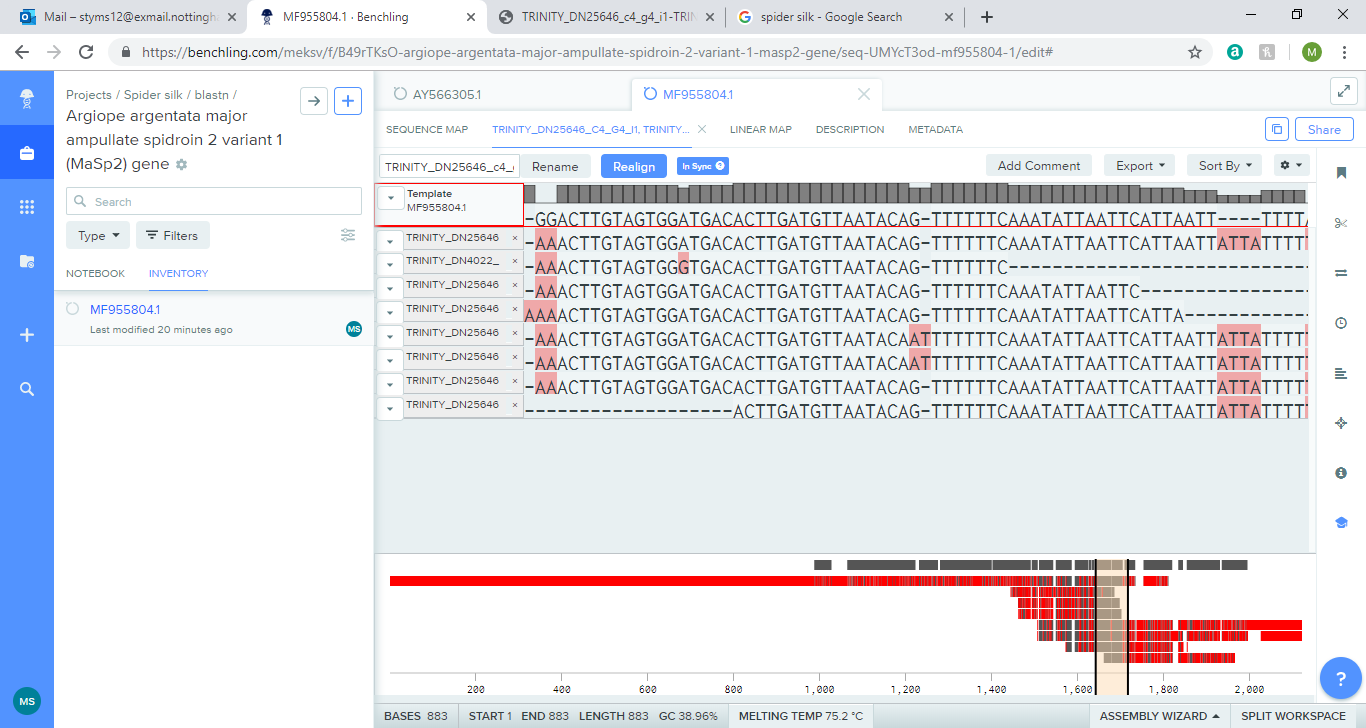
**3.2.2 Argiope argentata major ampullate spidroin 2 variant 1 (MaSp2) gene**

**Figure 11:** Sequence alignment for Argiope argentata major ampullate spidroin 2 variant 1 (MaSp2) gene. Red lines mean different nucleotide. Spaces between red lines means same nucleotide.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Query ID** | **Number of gap openings** | **Start of alignment in query** | **End of alignement in query** | **Start of alignment in subject** | **End of alignment in subject** | **Query frame** |
| TRINITY\_DN25646\_c4\_g4\_i1 | 2 | 53 | 236 | 650 | 467 | 1 |
| TRINITY\_DN4022\_c0\_g1\_i1 | 1 | 1 | 75 | 624 | 551 | 1 |
| TRINITY\_DN25646\_c4\_g2\_i1 | 1 | 1 | 87 | 636 | 551 | 1 |
| TRINITY\_DN25646\_c4\_g7\_i1 | 1 | 1 | 91 | 640 | 551 | 1 |
| TRINITY\_DN25646\_c4\_g3\_i2 | 3 | 376 | 560 | 650 | 467 | 1 |
| TRINITY\_DN25646\_c4\_g3\_i4 | 3 | 364 | 548 | 650 | 467 | 1 |
| TRINITY\_DN25646\_c4\_g3\_i1 | 2 | 122 | 248 | 650 | 524 | 1 |
| TRINITY\_DN25646\_c4\_g3\_i3 | 1 | 231 | 279 | 650 | 601 | 1 |

**Table 7:** Blastn for paired sequencing for Argiope argentata major ampullate spidroin 2 variant 1 (MaSp2) gene. Some query IDs have multiple rows in blastn document that differ in final 5 column headings only. Out of those multiple rows, the data shown is the row that is ordered first

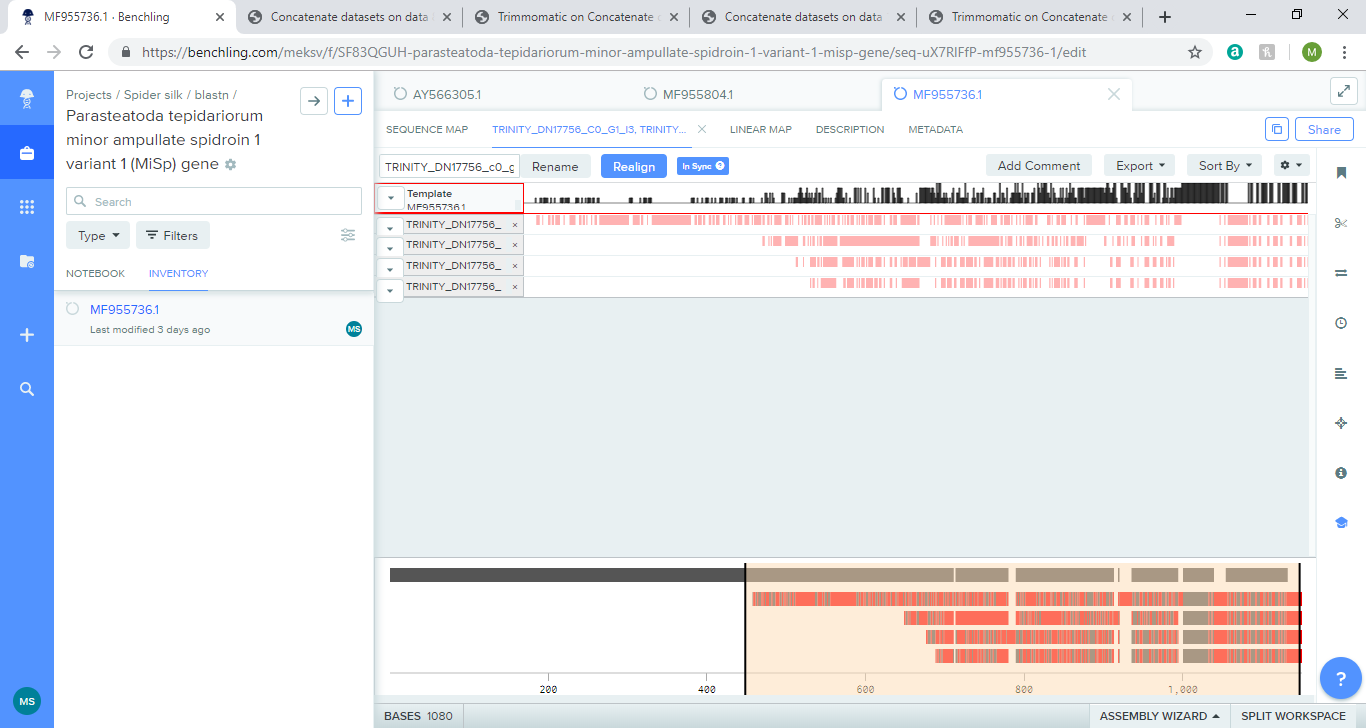
Compared to previous sequence (3.3.1), there is an even higher percentage of identical matches ranging from 85.3%-96.0% (3sf/1dp), so the sequences are very similar. Sequences are translated in forward direction (5’ to 3’), due to positive number and start with 1st reading frame. Based on the start and end of alignment in subject values, the sequence is most likely reverse complimented. A section of the template sequence aligns well with many of the query sequences, starting 620-650 bp from the subject. This suggests that this is a conserved sequence between spider species and if that sequence changed the protein translated by this mRNA may be non-functional. A better visual of this alignment is provided below.



**Figure 12:** Conserved section of sequence alignment for Argiope argentata major ampullate spidroin 2 variant 1 (MaSp2) gene. Red means different nucleotide. The particular section of alignment is highlighted in light orange on bottom half of image.

**3.2.3 Parasteatoda tepidariorum minor ampullate spidroin 1 variant 1 (MiSp) gene**

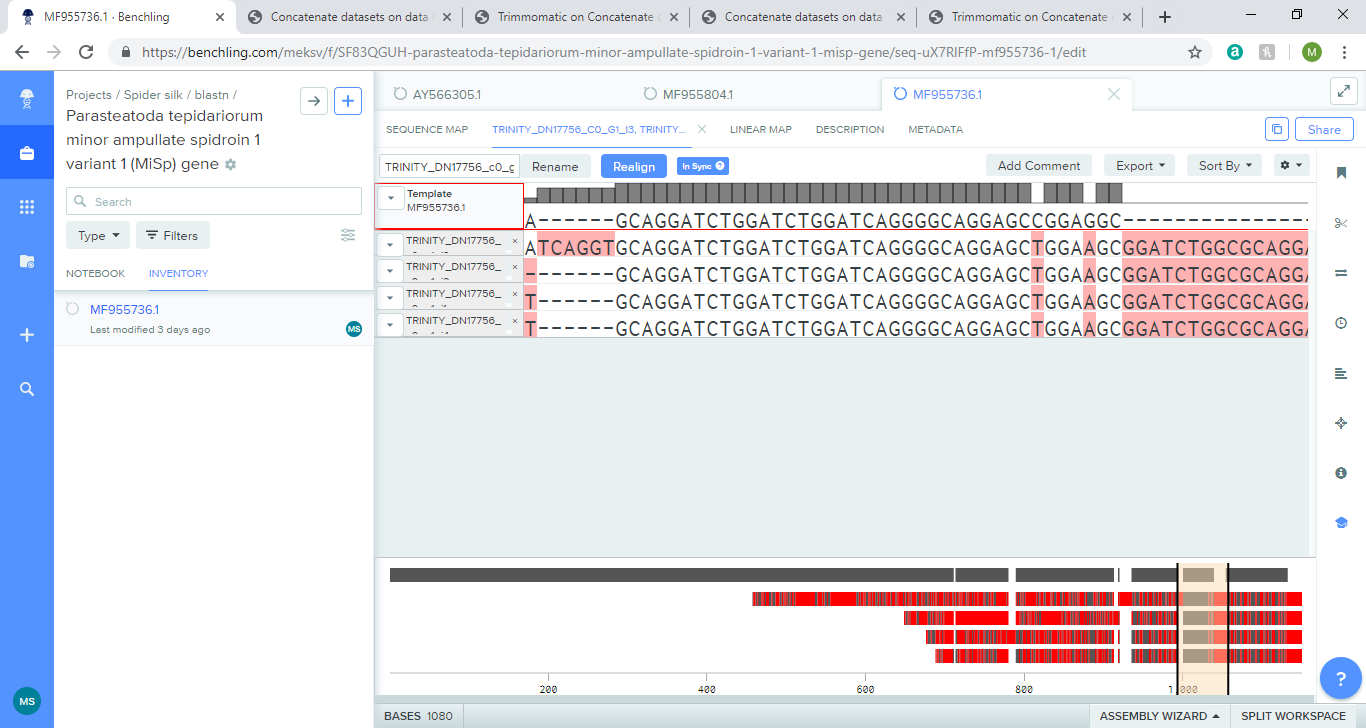
**Table 5:** Blastn for paired sequencing for Parasteatoda tepidariorum minor ampullate spidroin 1 variant 1 (MiSp) gene. Some query IDs have multiple rows in blastn document that differ in final 5 column headings only. Out of those multiple rows, the data shown is the row that is ordered first

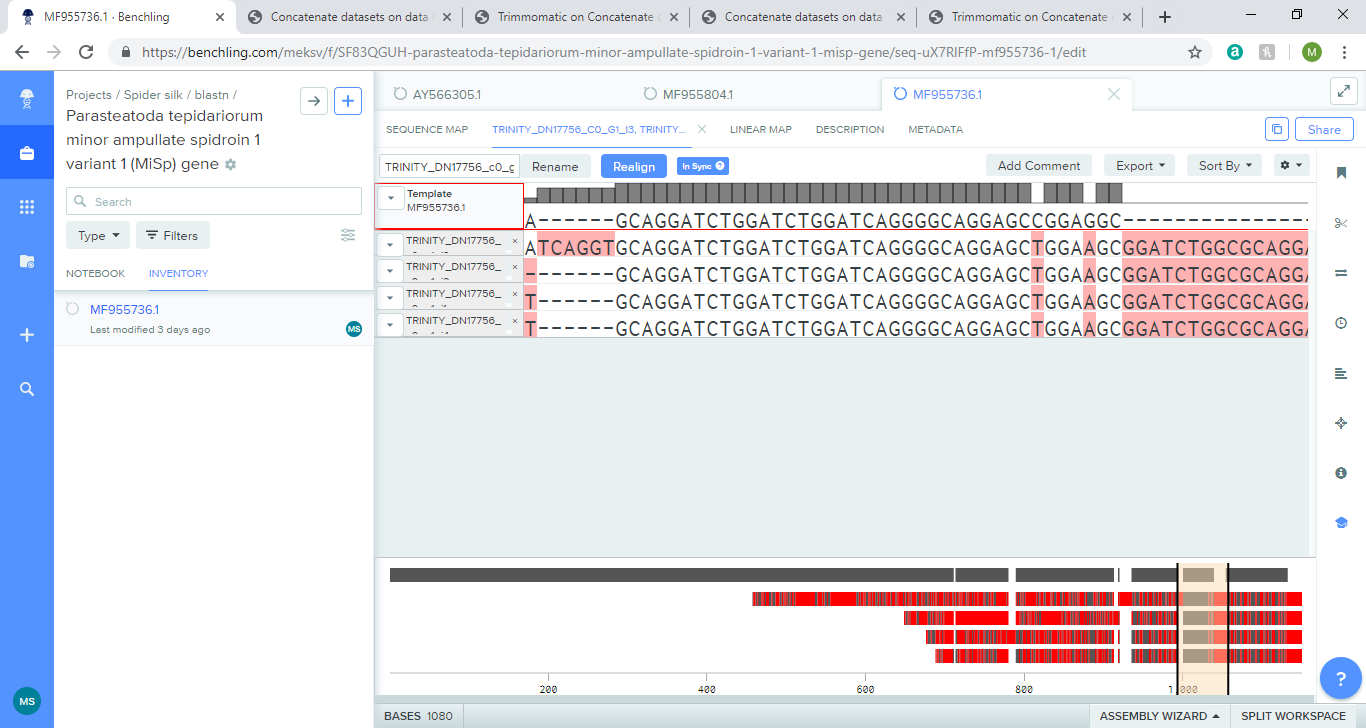


**Figure 13:** Sequence alignment for Parasteatoda tepidariorum minor ampullate spidroin 1 variant 1 (MiSp) gene. Red lines mean different nucleotide. Spaces between red lines means same nucleotide.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Query ID** | **Number of gap openings** | **Start of alignment in query** | **End of alignement in query** | **Start of alignment in subject** | **End of alignment in subject** | **Query frame** |
| TRINITY\_DN17756\_c0\_g1\_i3 | 0 | 245 | 331 | 786 | 872 | 1 |
| TRINITY\_DN17756\_c0\_g1\_i2 | 0 | 192 | 230 | 961 | 999 | 1 |
| TRINITY\_DN17756\_c0\_g1\_i1 | 5 | 245 | 379 | 786 | 920 | 1 |
| TRINITY\_DN17756\_c0\_g1\_i4 | 5 | 209 | 343 | 786 | 920 | 1 |

A high percentage of identical matches ranging from 78.7% to 95.3% (3sf), means that there is a high level of sequence similarity. All aligned sequences are in query frame 1. Sequences are translated in forward direction (5’ to 3’), due to positive number and start with 1st reading frame. Four of the six sequences begin their alignment with the subject at 786bp and ending at 872bp, with one sequence starting at 787bp instead (and ending at 872bp, as well). All aligned sequences are isoforms of each other as the query id is identical, except for the number after ‘i‘. Similar to 3.3.2, there is a section of the template sequence that aligns well with many of the query sequences, around 1000 bp from the subject. This suggests that this is a conserved sequence between spider species and if that sequence changed the protein translated by this mRNA may be non-functional. A better visual of this alignment is provided below.



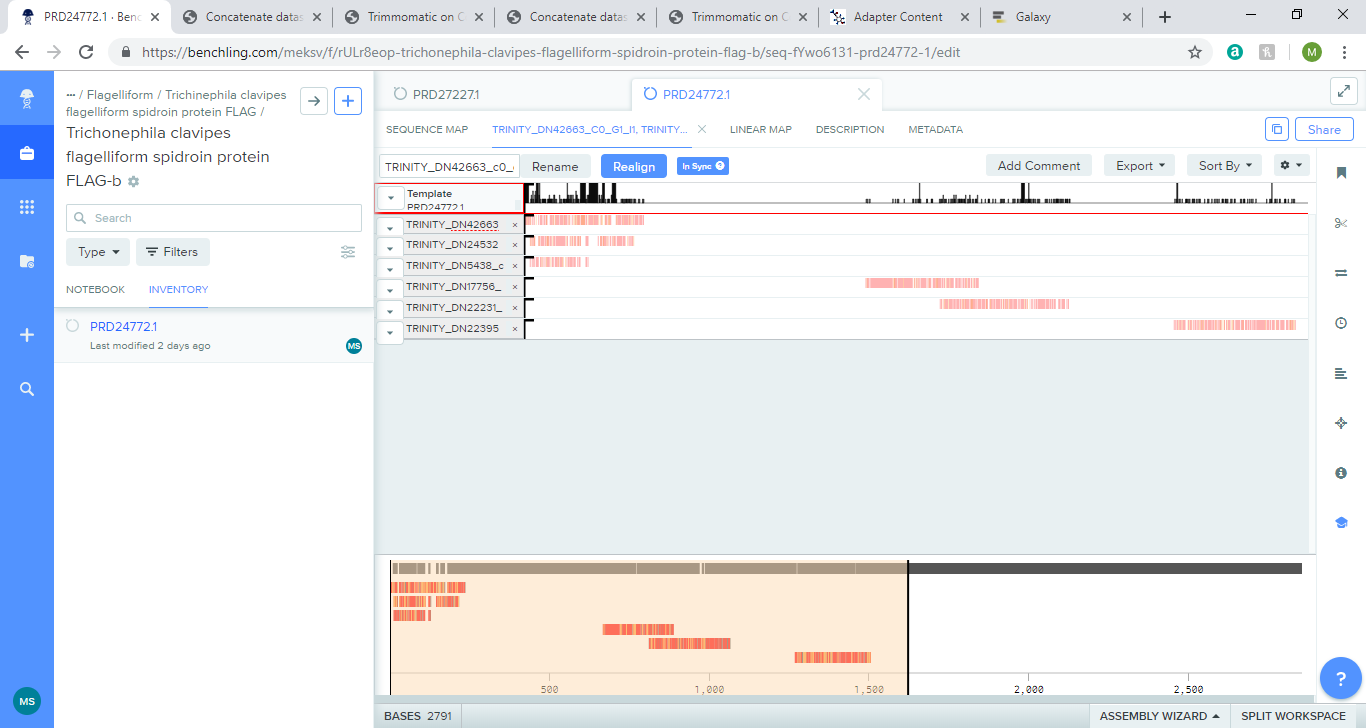
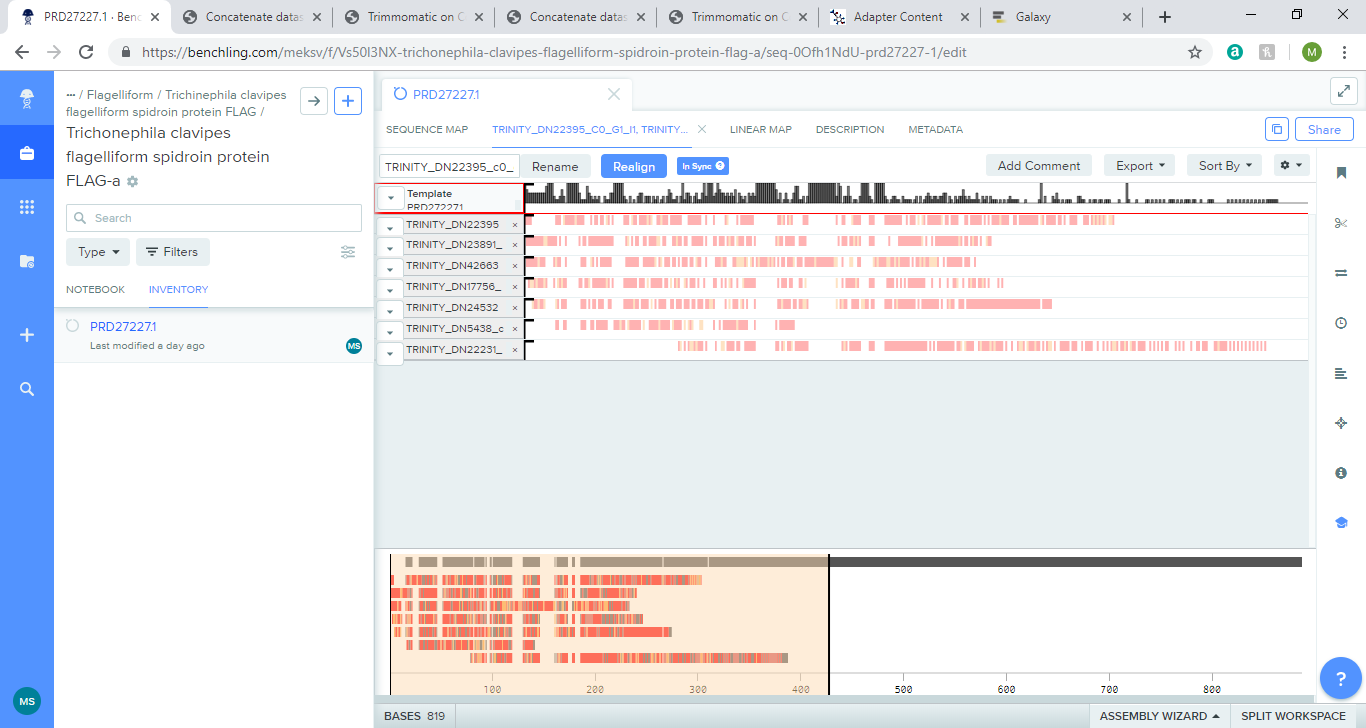


**Figure 14:** Conserved section of sequence alignment for Parasteatoda tepidariorum minor ampullate spidroin 1 variant 1 (MiSp) gene. Red means different nucleotide. The particular section of alignment is highlighted in light orange on bottom half of image.

**3.3 Blastp**

From the blastp output from BLAST+, x sets of proteins that are directly related to silk have been taken from NCBI and sequence aligned to various query sequences from the AssemblyPostProcessor tool on Galaxy. The sequence alignment will be shown but to gather more data the amino acid composition will be compared to sequences, if they fall into certain regions.

**3.3.1 Flagelliform**

**3.3.1.1 Trichonephila clavipes flagelliform spidroin protein FLAG**

**Figure 15:** Sequence alignment for Trichonephila clavipes flagelliform spidroin protein FLAG-a and FLAG-b. Red lines mean different nucleotide. Spaces between red lines means same nucleotide.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Query ID** | **E value** | **Bit score** | **Percentage of identical matches** | **Alignment length** | **Number of mismatches** | **Number of gap openings** |
| **FLAG-a** |  |  |  |  |  |  |
| TRINITY\_DN22395\_c0\_g1\_i1 | 2.86E-14 | 67.4 | 36.066 | 122 | 75 | 1 |
| TRINITY\_DN23891\_c1\_g3\_i1 | 3.43E-06 | 42.4 | 22.656 | 128 | 98 | 1 |
| TRINITY\_DN42663\_c0\_g1\_i1 | 1.45E-08 | 50.1 | 30.556 | 72 | 50 | 0 |
| TRINITY\_DN17756\_c0\_g2\_i2 | 1.99E-18 | 78.2 | 30.469 | 128 | 88 | 1 |
| TRINITY\_DN24532\_c2\_g1\_i1 | 7.18E-19 | 79.7 | 30.818 | 159 | 104 | 2 |
| TRINITY\_DN5438\_c1\_g1\_i1 | 2.15E-13 | 61.2 | 35.036 | 137 | 81 | 5 |
| TRINITY\_DN22231\_c1\_g1\_i1 | 1.13E-19 | 83.6 | 32.331 | 133 | 88 | 2 |
| **FLAG-b** |  |  |  |  |  |  |
| TRINITY\_DN24532\_c2\_g1\_i1 | 8.13E-16 | 70.9 | 29.927 | 137 | 90 | 2 |
| TRINITY\_DN17756\_c0\_g2\_i2 | 1.31E-15 | 70.1 | 24.161 | 149 | 105 | 2 |
| TRINITY\_DN5438\_c1\_g1\_i1 | 1.68E-15 | 67.4 | 36.923 | 130 | 78 | 3 |

**Table 6:** Blastp of Trichonephila clavipes flagelliform spidroin protein FLAG. Top table is the 5 sequences in order of aligned sequences from sequence alignment for FLAG-a. Bottom table is the 3 sequences in order of aligned sequences from sequence alignment for FLAG-b (only in possible N terminus region).

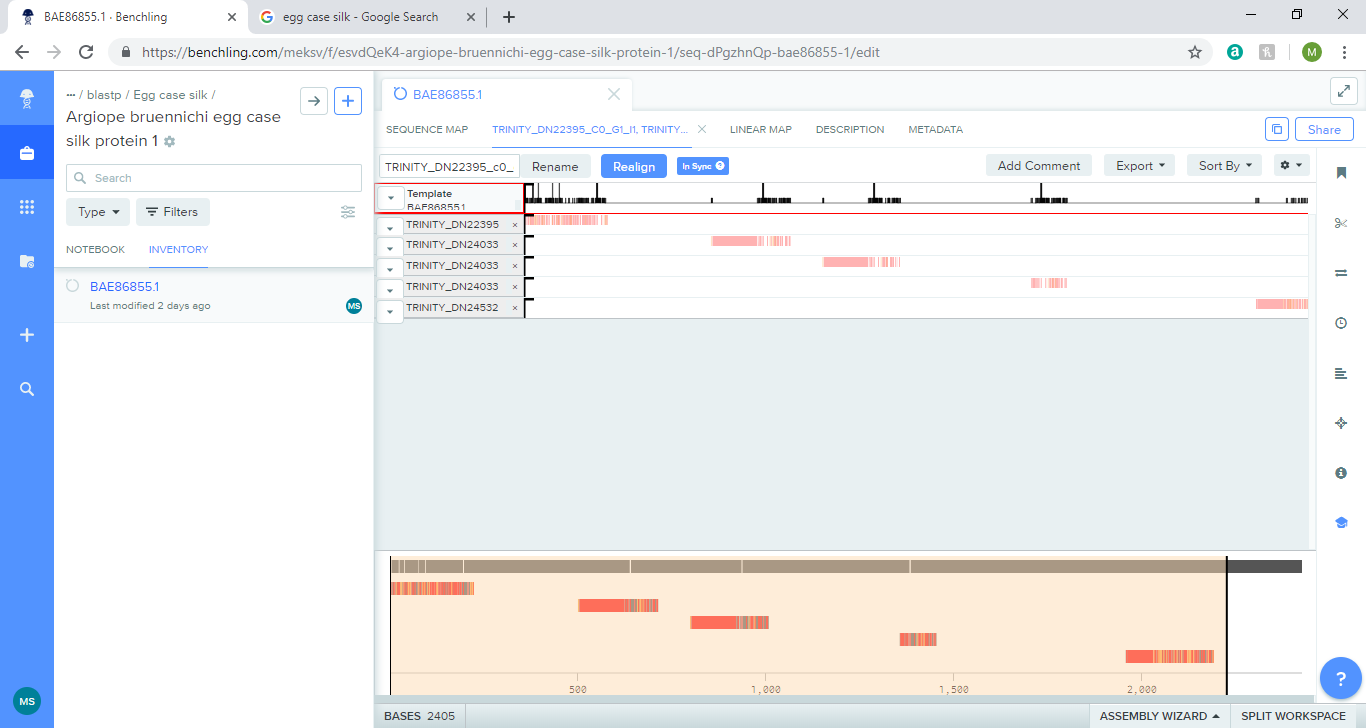
Unlike the nucleotide sequences, the protein sequences have not aligned very well as the percentage of identical matches ranges from 30.0 to 37.0 % (3sf). The number of mismatches is quite high given the corresponding alignment length. The query sequences align to the first half of the protein, so the amino acid composition of the N terminus of the template is used for comparison to the first 160bp of the query sequences. Broadly speaking, there seems to be a consistent amino acid composition between the FLAG and aligned sequences.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | FLAG A Spidroin\_N | FLAG B Spidroin\_N | 22395  c0g1i1 | 23891  c1g3i1 | 42663  c0g1i1 | 17756  c0g2i2 | 24532  c2g1i1 | 5438  c1g1i1 | 22231  c1g1i1 |
| Ala | 12.1 | 11.5 | 14.4 | 7.5 | 8.8 | 7.5 | 6.9 | 9.0 | 10.0 |
| Gly | 6.5 | 6.6 | 8.8 | 4.4 | 5.0 | 5.6 | 3.1 | 2.8 | 6.9 |
| Ser | 12.1 | 10.7 | 10.0 | 20.6 | 9.4 | 13.8 | 11.2 | 12.5 | 15.0 |
| Gln | 8.1 | 7.4 | 6.2 | 5.0 | 3.1 | 8.1 | 6.2 | 5.6 | 7.5 |
| Glu | 8.1 | 7.4 | 4.4 | 7.5 | 12.5 | 7.5 | 6.9 | 9.0 | 3.1 |
| Gly | 6.5 | 6.6 | 8.8 | 4.4 | 5.0 | 5.6 | 3.1 | 2.8 | 6.9 |
| Ile | 7.3 | 8.2 | 7.5 | 3.8 | 9.4 | 7.5 | 8.8 | 5.6 | 8.8 |
| Leu | 5.6 | 4.9 | 11.2 | 3.1 | 9.4 | 6.2 | 8.1 | 8.3 | 5.6 |
| Ser | 12.1 | 10.7 | 10.0 | 20.6 | 9.4 | 13.8 | 11.2 | 12.5 | 15.0 |
| Thr | 4.0 | 4.9 | 7.5 | 5.0 | 3.8 | 6.9 | 9.4 | 6.9 | 5.6 |
| Val | 6.5 | 6.6 | 6.9 | 7.5 | 10.0 | 4.4 | 5.6 | 6.9 | 5.0 |

**Table 8:** Amino acid composition (%) of Trichonephila clavipes flagelliform spidroin protein. Yellow box means equal or more than 5%. Orange box means equal or more than 10%.

Table 8 shows certain amino acid compositions of the template sequences and the aligned sequences where there is a slight difference. Ser is very high in most sequences; Ala is high in most sequences and Gly has a lower percentage than both of those amino acids but is still high. In some cases, there is a slight difference such as TRINITY\_DN22395\_c1g3i1 having a higher percentage of Ser than template sequences. TRINITY\_DN24532\_c2g1i1 and TRINITY\_DN5438\_c1g1i1 have less than half of Gly compared to the template sequences. Apart from those three typical amino acids, each sequence has at least one amino acid that has slightly less or more percentages that the template sequence and this is visualised as the white box in each row in Table 7. All other amino acids had a similar composition and can be found in the amino acid composition excel file.

**3.3.2 Egg case silk**

**3.3.2.1 Argiope bruennichi egg case silk protein**

**Figure 16**: Sequence alignment for Argiope bruennichi egg case silk protein 1. Red lines mean different nucleotide. Spaces between red lines means same nucleotide.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Query ID** | **E value** | **Bit score** | **Percentage of identical matches** | **Alignment length** | **Number of mismatches** | **Number of gap openings** |
| TRINITY\_DN22395\_c0\_g1\_i1 | 5.39E-45 | 157 | 48.763 | 283 | 124 | 4 |
| TRINITY\_DN24532\_c2\_g1\_i1 | 4.49E-19 | 80.5 | 37.107 | 159 | 86 | 5 |

**Table 9:** Blastp of Argiope bruennichi egg case silk protein 1. The table shows data for the first and last query sequence that aligns with the template sequence.

TRINITY\_DN22395\_c0\_g1\_i1 seems to align at the N terminus and has the highest percentage of identical match for proteins seen so far with 48.8% but it is still below half. For both sequences, the number of mismatches is nearly half of the alignment length. TRINITY\_DN24532\_c2\_g1\_i1 may align to the C terminus given that it is nearer the end of the protein. Using the axis options on benchling, the alignment is between 2020-2170bp from the template sequence and 20bp-170bp from the query sequence.

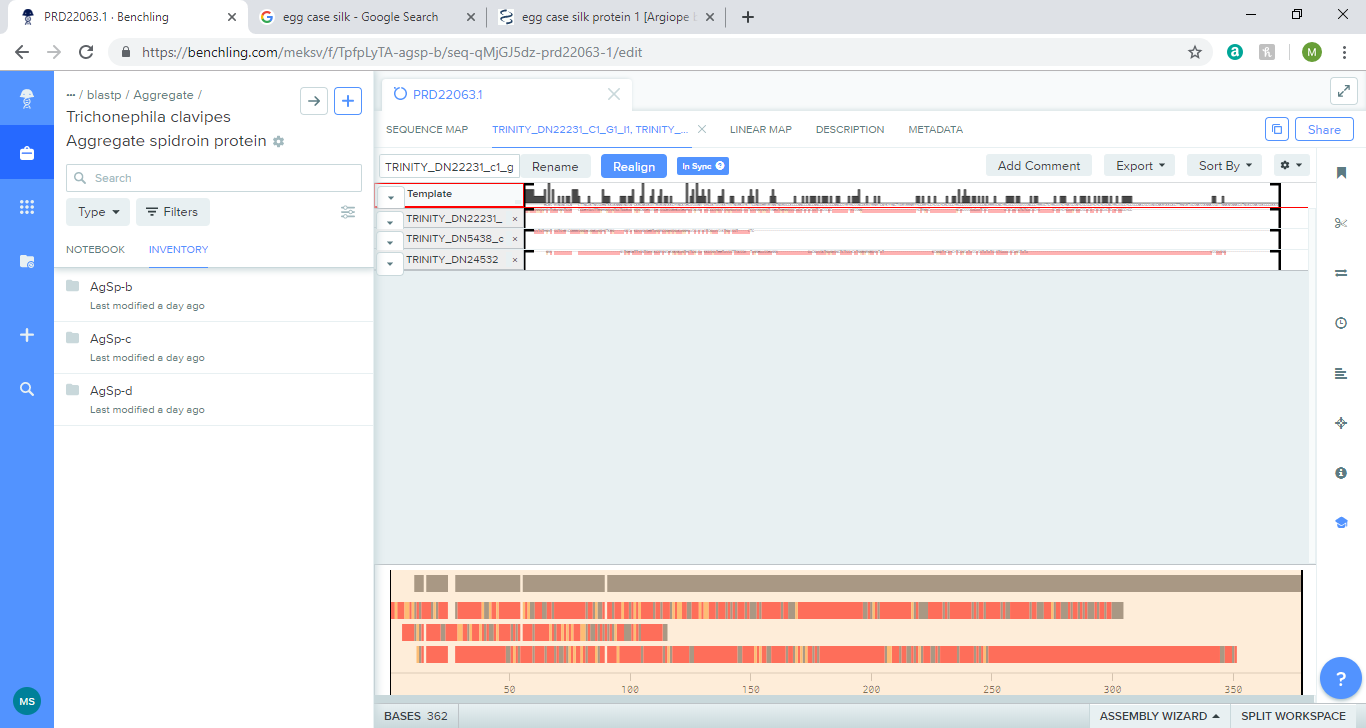
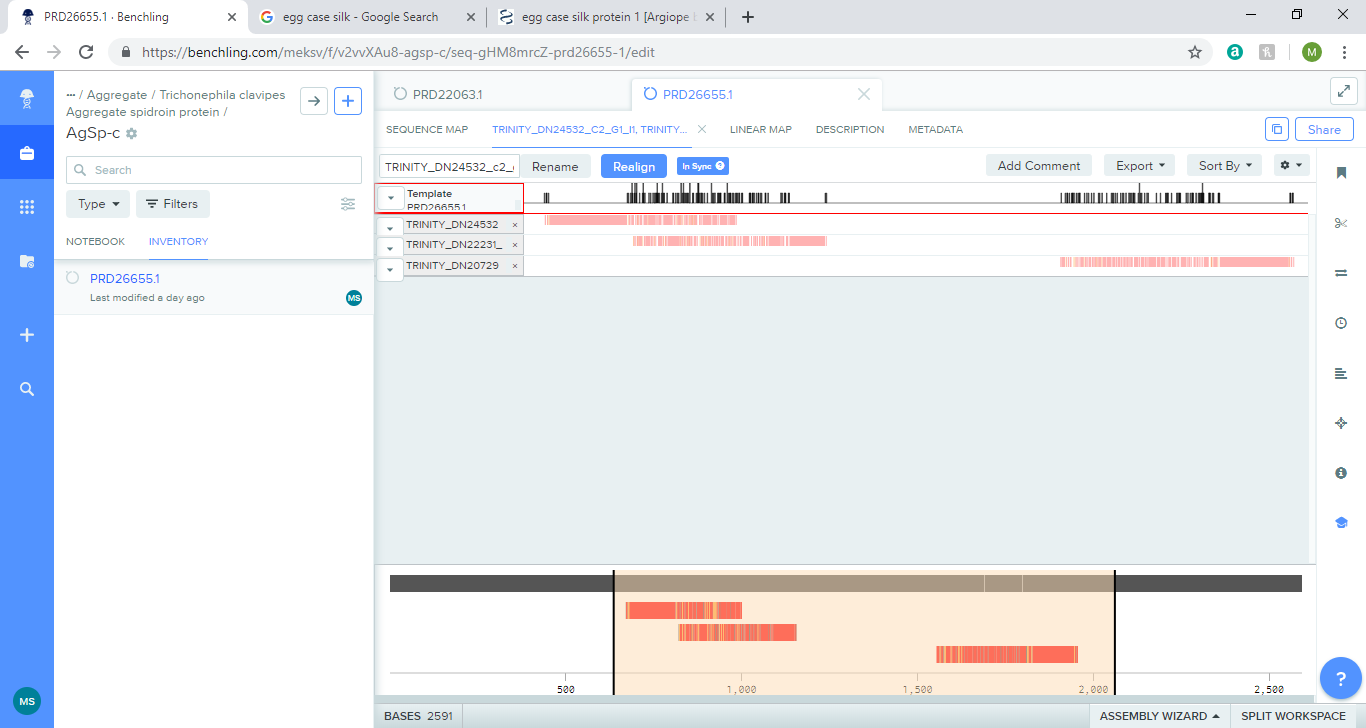
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Egg case silk, N-term | 22395\_c0\_g1\_i1 | Egg case silk, 2020-2170bp | 24532\_c2\_g1\_i1, 20-170bp |
| Ala | 13.2 | 14.4 | 27.2 | 7.9 |
| Gly | 5.0 | 8.8 | 10.6 | 4.0 |
| Ser | 16.5 | 10.0 | 23.2 | 11.3 |
| Glu | 3.3 | 4.4 | 0.0 | 9.3 |
| Ile | 6.6 | 7.5 | 0.7 | 6.6 |
| Lys | 0.0 | 1.2 | 0.0 | 6.0 |
| Thr | 7.4 | 7.5 | 4.6 | 9.3 |

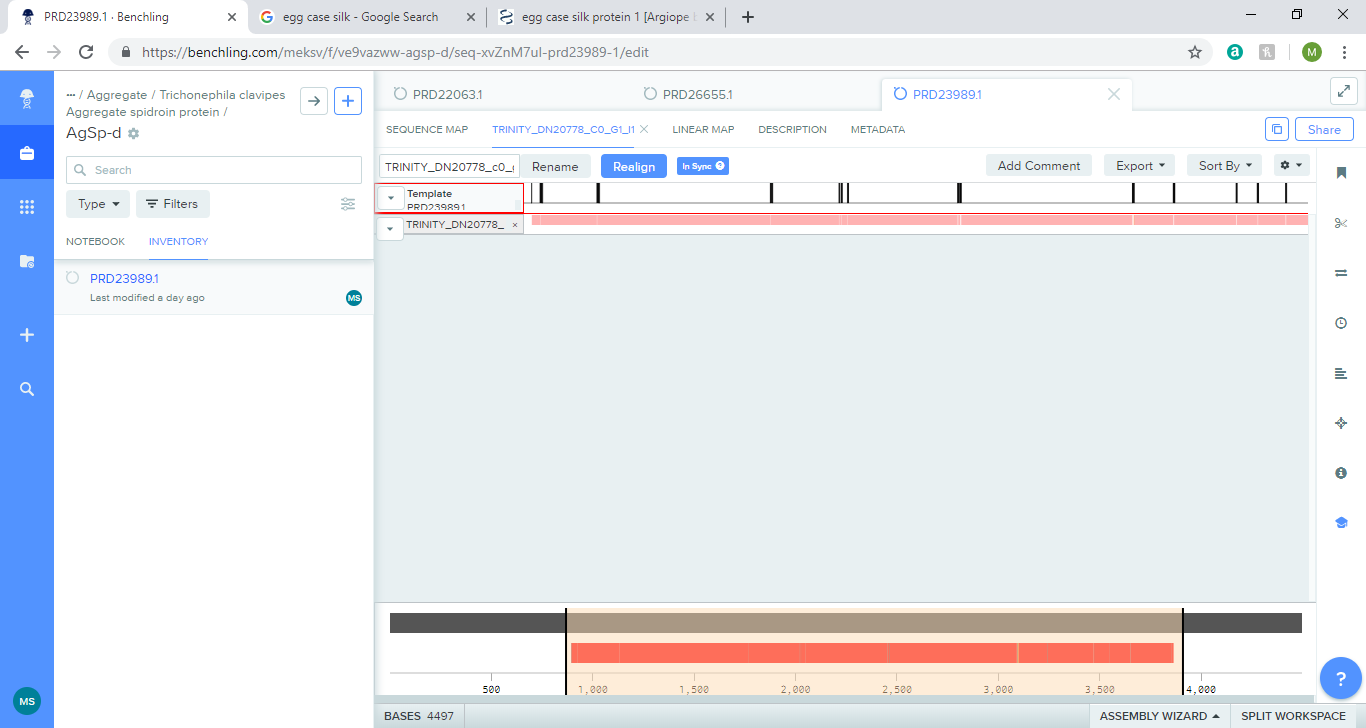
**Table 10:** Amino acid composition (%) of Argiope bruennichi egg case silk protein 1. Yellow box means equal or more than 5%. Orange box means equal or more than 10%.

Table 10 shows certain amino acid compositions of the template sequences and the aligned sequences where there is a slight difference. Ser is very high in most sequences; Ala is high in most sequences and Gly has a lower percentage than both of those amino acids but is still high. In general, there does not seem that be any major differences between the n-terminus and the egg case silk. However, in the 2020-2170bp region that aligns with the query sequence, Ala and Ser has a much higher composition (more than 10%) in the 2020-2170bp than in the query sequence, while the Glu, Ile and Lys have a higher composition (more than 5%) in the query sequence compared to the egg case silk. For Thr, there is a higher composition in the query sequence than the egg case silk but not more than a 5% difference. The difference in the amino acid compositions suggest that this may not be the C terminus and 2020-2170bp may be part of the repetitive region, as in general the C terminus is more conserved between spider species. All other amino acids had a similar composition and can be found in the amino acid composition excel file.

**3.3.3 Aggregate**

**3.3.3.1 Trichonephila clavipes Aggregate spidroin protein**





**Figure 17:** Sequence alignment of Trichonephila clavipes Aggregate spidroin protein AgSp-b (top), AgSp-c (middle), AgSp-d (bottom). Red lines mean different nucleotide. Spaces between red lines means same nucleotide.

**Table 5 and 6:** Blastp of Trichonephila clavipes Aggregate spidroin protein AgSp-b (top), AgSp-c, AgSp-d. The table shows the order of query sequences that aligns with the template sequence.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Query ID** | **E value** | **Bit score** | **Percentage of identical matches** | **Alignment length** | **Number of mismatches** | **Number of gap openings** |
| **AgSp-b** |  |  |  |  |  |  |
| TRINITY\_DN22231\_c1\_g1\_i1 | 4.74E-07 | 45.4 | 22.353 | 170 | 124 | 4 |
| TRINITY\_DN5438\_c1\_g1\_i1 | 6.91E-07 | 42.4 | 24.793 | 121 | 87 | 3 |
| TRINITY\_DN24532\_c2\_g1\_i1 | 4.34E-08 | 47.8 | 26.786 | 112 | 78 | 3 |
| **AgSp-c** |  |  |  |  |  |  |
| TRINITY\_DN24532\_c2\_g1\_i1 | 2.42E-07 | 45.8 | 26.389 | 144 | 103 | 2 |
| TRINITY\_DN22231\_c1\_g1\_i1 | 5.27E-13 | 63.9 | 25.625 | 160 | 104 | 2 |
| TRINITY\_DN20729\_c0\_g1\_i5 | 7.41E-06 | 42.4 | 26.786 | 168 | 115 | 3 |
| **AgSp-d** |  |  |  |  |  |  |
| TRINITY\_DN20778\_c0\_g1\_i1 | 3.42E-09 | 47.4 | 38.318 | 107 | 58 | 5 |

**Table 11:** Blastp of Trichonephila clavipes Aggregate spidroin protein.The table shows data for the first and last query sequence that aligns with the template sequence.

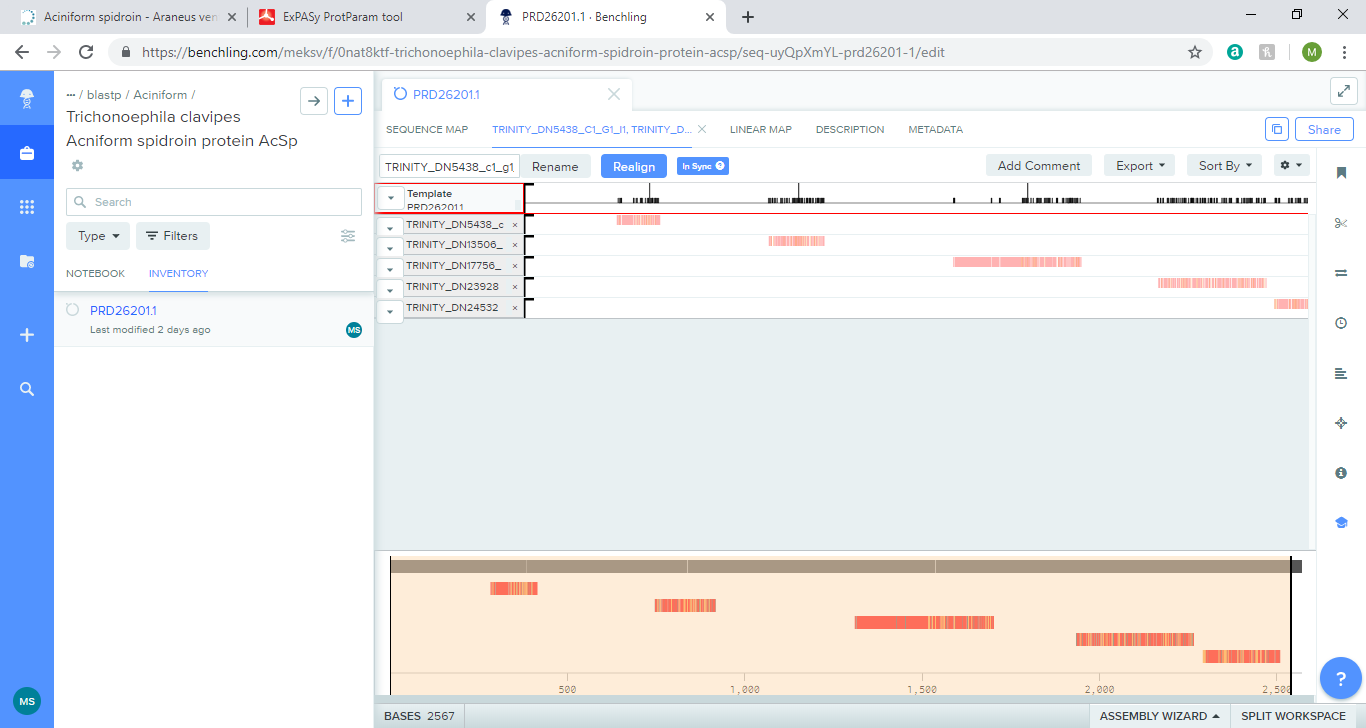
The first three aligned sequences of AgSp-b fall in the N terminus region. The percentage of identical matches is less than half and the number of mismatches compared to alignment length suggests that the sequence alignment is very poor.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AgSp-b,  N-term | 22231  c1g1i1, 30-150bp | 5438  c1g1i1, 30-150bp | 24532  c2g1i1, 30-150bp |
| Ala | 8.7 | 11.6 | 11.3 | 9.1 |
| Gly | 6.1 | 6.6 | 2.6 | 4.1 |
| Ser | 4.3 | 17.4 | 12.2 | 13.2 |
| Gln | 1.7 | 8.3 | 5.2 | 5.0 |
| Leu | 12.2 | 3.3 | 6.1 | 5.0 |
| Lys | 12.2 | 1.7 | 4.3 | 6.6 |

**Table 12:** Amino acid composition (%) of Trichonephila clavipes Aggregate spidroin protein. Yellow box means equal or more than 5%. Orange box means equal or more than 10%.

Table 12 shows certain amino acid compositions of the template sequences and the aligned sequences where there is a slight difference. Unusually, Ser is quite low in the template sequence compared to other proteins, but the template sequences have a typical high Ser. In most sequences; Ala is high in most sequences and Gly has a lower percentage than both of those amino acids with TRINITY\_DN5438\_c1g1i1 having quite a lower percentage. In some cases, there is a slight difference such as AgSp-b having a much lower than amino acid composition for Gln than the query sequences, while the Leu and Lys are much higher. All other amino acids had a similar composition and can be found in the amino acid composition excel file.

**3.3.4 Aciniform**

**3.3.4.1 Trichonoephila clavipes Acniform spidroin protein AcSp**

**Figure 18:** Sequence alignment of Trichonephila clavipes Acniform spidroin protein AcSp. Red lines mean different nucleotide. Spaces between red lines means same nucleotide.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Query ID** | **E value** | **Bit score** | **Percentage of identical matches** | **Alignment length** | **Number of mismatches** | **Number of gap openings** |
| TRINITY\_DN23928\_c0\_g1\_i3 | 3.61E-07 | 47 | 25 | 300 | 196 | 4 |
| TRINITY\_DN24532\_c2\_g1\_i1 | 1.24E-15 | 70.5 | 28.729 | 181 | 101 | 6 |

**Table 13:** Blastp of Trichonephila clavipes Acniform spidroin protein AcSp.The table shows data for the first and last query sequence that aligns with the template sequence.

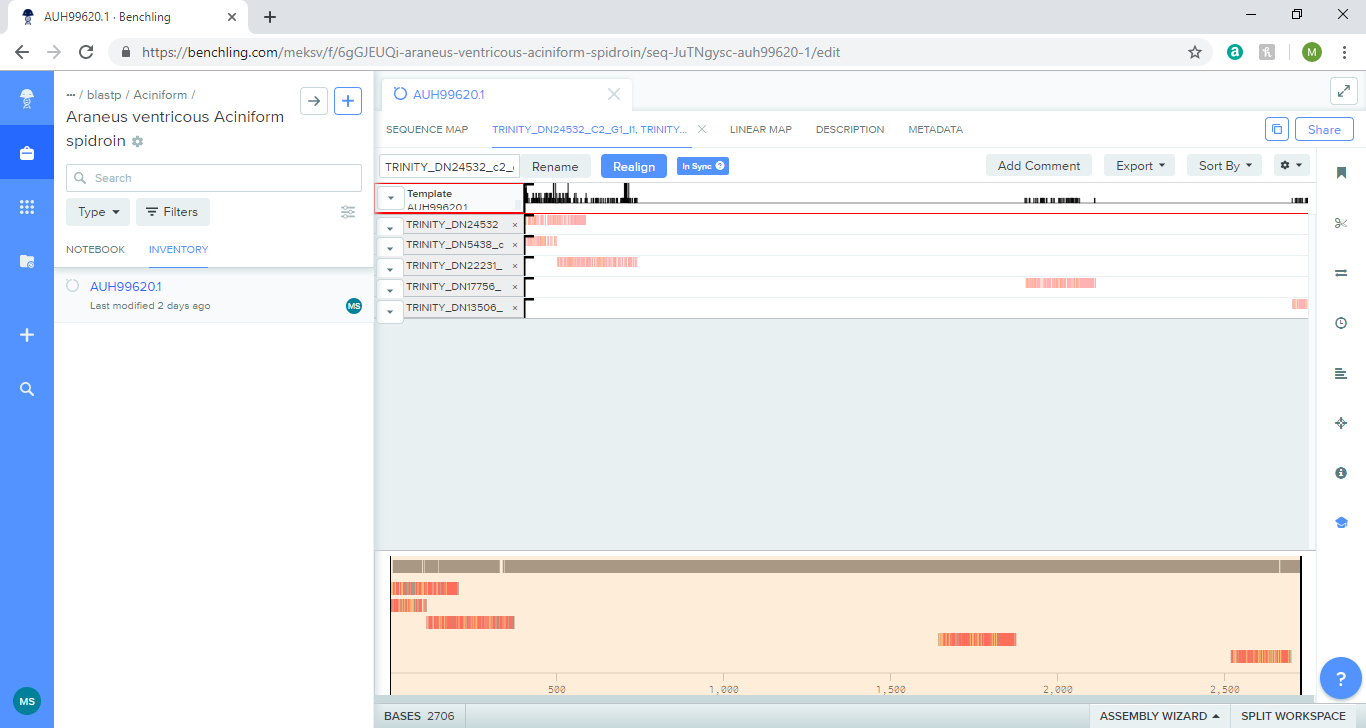
The last two aligned sequences might fall into the C terminus of the protein. Both sequences are very poorly aligned as the number of mismatches is more than half the alignment length and the percentage of identical match is less than 30%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | AcSp, 1950-2250bp | 23928  c0\_g1\_i3 | AcSp,  2137-2463bp | 24532  c2\_g1\_i1 |
| Ala | 12.0 | 13.8 | 13.6 | 11.1 |
| Gly | 10.3 | 8.8 | 5.4 | 4.4 |
| Ser | 17.3 | 30.9 | 17.7 | 15.0 |
| Val | 10.0 | 4.4 | 11.6 | 4.9 |
| Glu | 1.7 | 0.8 | 1.4 | 6.6 |

**Table 14:** Amino acid composition (%) of Trichonephila clavipes Acniform spidroin protein AcSp. Yellow box means equal or more than 5%. Orange box means equal or more than 10%.

Table 14 shows certain amino acid compositions of the template sequences and the aligned sequences where there is a slight difference. Most of the amino acids had a similar composition and can be found in the amino acid composition excel file, including Ala, Gly and Ser. The percentage of Val is high in both regions of AcSp, but low in the query sequences. The percentage of Glu is higher in TRINITY\_24532\_c2\_g1\_i1 compared to the corresponding AcSp region (more than 5%). Given the high similarity in amino composition in the 2137-2463bp region of AcSp and the query sequence, this suggests that this region is in the C-terminus and not the repetitive region.

**3.3.4.2 Araneus ventricous Aciniform spidroin**



**Figure 19:** Sequence alignment of Araneus ventricous Aciniform spidroin. Red lines mean different nucleotide. Spaces between red lines means same nucleotide.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Query ID** | **E value** | **Bit score** | **Percentage of identical matches** | **Alignment length** | **Number of mismatches** | **Number of gap openings** |
| TRINITY\_DN24532\_c2\_g1\_i1 | 7.79E-41 | 143 | 46.154 | 156 | 82 | 2 |
| TRINITY\_DN5438\_c1\_g1\_i1 | 4.26E-37 | 129 | 48.551 | 138 | 69 | 2 |
| TRINITY\_DN22231\_c1\_g1\_i1 | 8.75E-21 | 87 | 29.091 | 165 | 106 | 3 |
| TRINITY\_DN13506\_c0\_g1\_i1 | 1.88E-13 | 63.2 | 39.073 | 151 | 86 | 3 |

**Table 15:** Blastp of Araneus ventricous Aciniform spidroin. The table shows data for the first and last query sequence that aligns with the template sequence.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Aciniform spidroin,N term | 24532  c2\_g1\_i1 | 5438  c1\_g1\_i1 | 22231  c1\_g1\_i1 | Aciniform  Spidroin, 2499-2641bp | 13506  c0\_g1\_i1 |
| Ala | 16.7 | 11.1 | 9.1 | 11.9 | 13.3 | 13.4 |
| Gly | 4.8 | 4.4 | 2.8 | 19.4 | 7.0 | 10.8 |
| Ser | 11.1 | 15.0 | 12.6 | 18.0 | 21.0 | 28.4 |
| Phe | 2.4 | 3.5 | 7.7 | 1.8 | 1.4 | 2.6 |
| Thr | 2.4 | 11.5 | 7.0 | 5.4 | 7.7 | 4.6 |

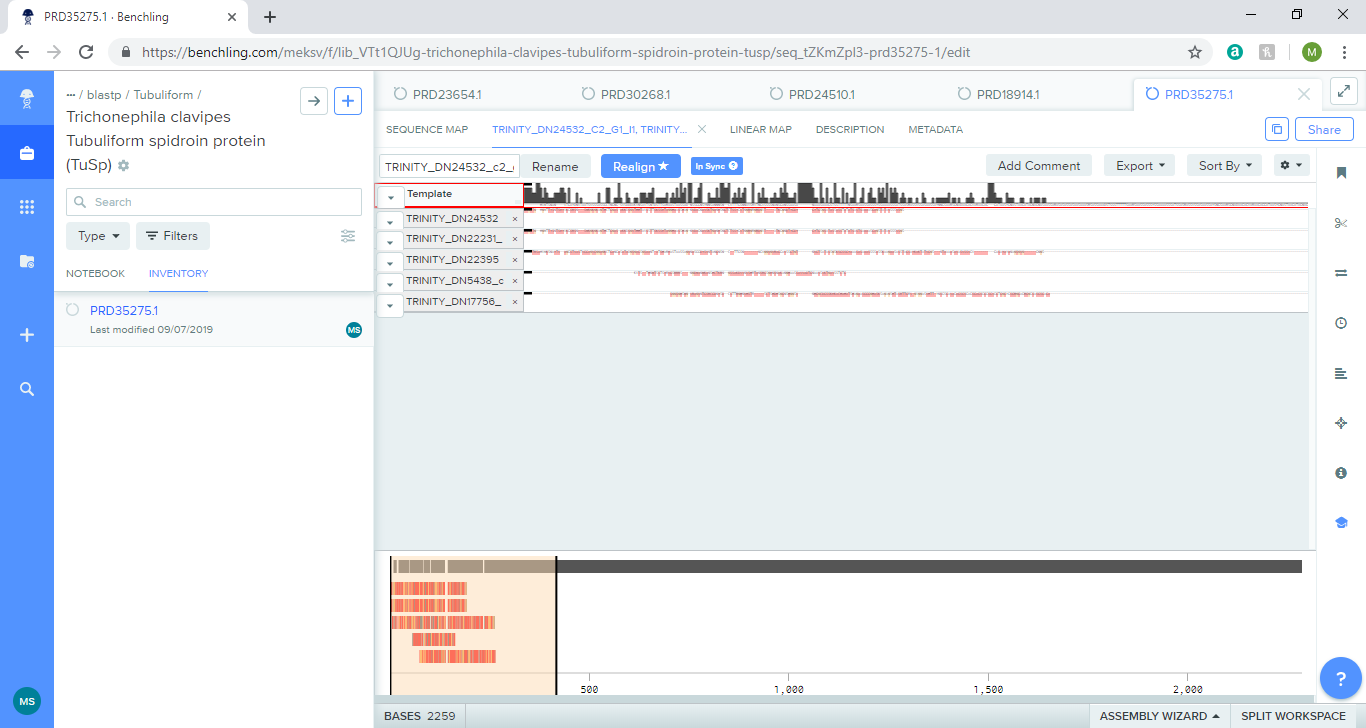
The first three sequences may fall in the N terminus region, while the last aligned sequence may fall in the C terminus region. The percentage of identical matches is less than half and the number of mismatches compared to alignment length suggests that the sequence alignment is very poor. TRINITY\_DN22231\_c1\_g1\_i1 is particularly poor with less than 30% of identical matches.

**Table 16:** Amino acid composition (%) of Araneus ventricous Aciniform spidroin. Yellow box means equal or more than 5%. Orange box means equal or more than 10%.

Table 16 shows certain amino acid compositions of the template sequences and the aligned sequences where there is a slight difference. Ser is very high in most sequences; Ala is high in most sequences except it is slightly lower in TRINTY\_DN5438\_c1\_g1\_i1. Gly has a lower percentage than both of those amino acids, except TRINITY\_DN22231\_c1\_g1\_i1 is a much higher percentage than the template sequence. Gly is higher in the 2499-2641bp region than the N terminus. Phe is higher in the TRINITY\_DN5438\_c1\_g1\_i1 than the template seqeucne , while Thr is higher in the query sequences than the N-term.

**3.3.5 Tubuliform**

**3.3.5.1 Trichonephila clavipes Tubuliform spidroin protein (TuSp)**



**Figure 20:** Sequence alignment of Trichonephila clavipes Tubuliform spidroin protein TuSp. Red lines mean different nucleotide. Spaces between red lines means same nucleotide.

+

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Query ID** | **E value** | **Bit score** | **Percentage of identical matches** | **Alignment length** | **Number of mismatches** | **Number of gap openings** |
| TRINITY\_DN24532\_c2\_g1\_i1 | 5.14E-25 | 97.8 | 33.548 | 155 | 93 | 4 |
| TRINITY\_DN22231\_c1\_g1\_i1 | 1.19E-16 | 74.7 | 29.677 | 155 | 105 | 2 |
| TRINITY\_DN22395\_c0\_g1\_i1 | 8.19E-43 | 151 | 55.102 | 147 | 58 | 2 |
| TRINITY\_DN5438\_c1\_g1\_i1 | 4.89E-15 | 65.9 | 31.884 | 138 | 84 | 3 |
| TRINITY\_DN17756\_c0\_g2\_i2 | 4.83E-15 | 68.6 | 26.708 | 161 | 111 | 3 |

**Table 17:** Blastp of Trichonephila clavipes Tubuliform spidroin protein TuSp. The table shows data for the first and last query sequence that aligns with the template sequence.

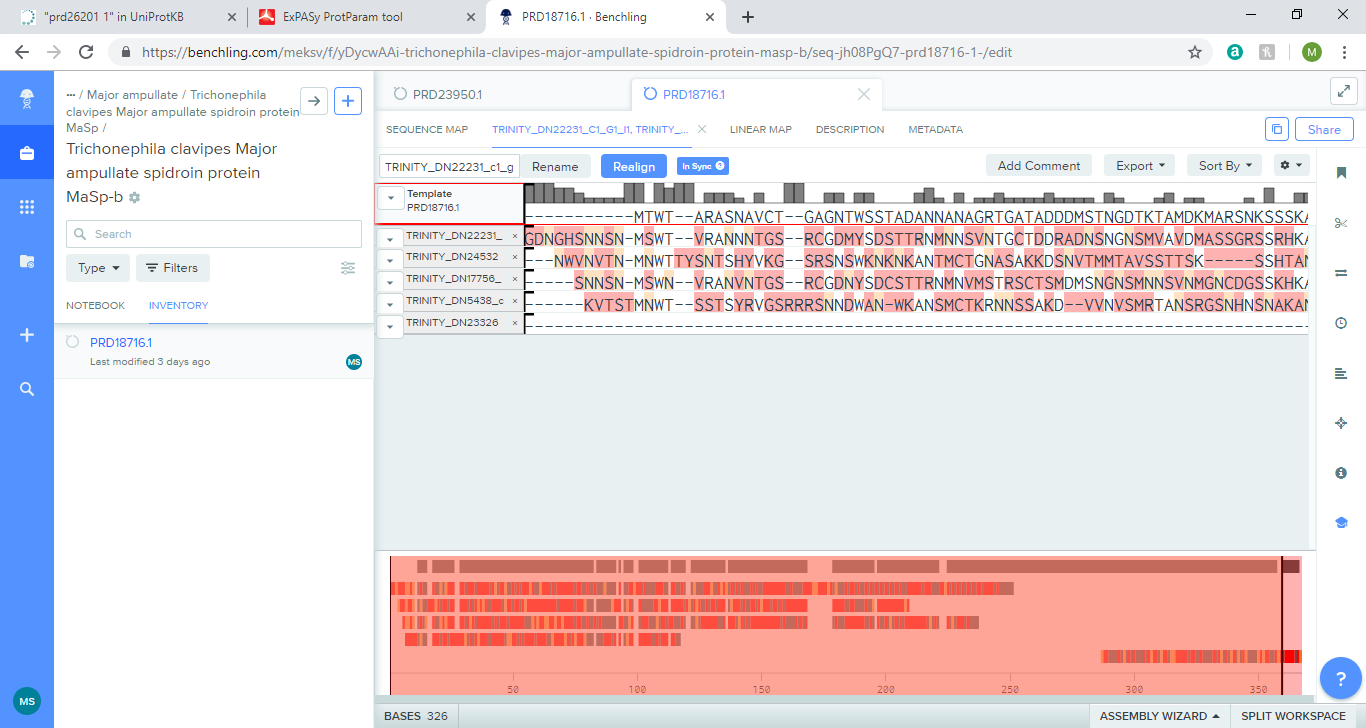
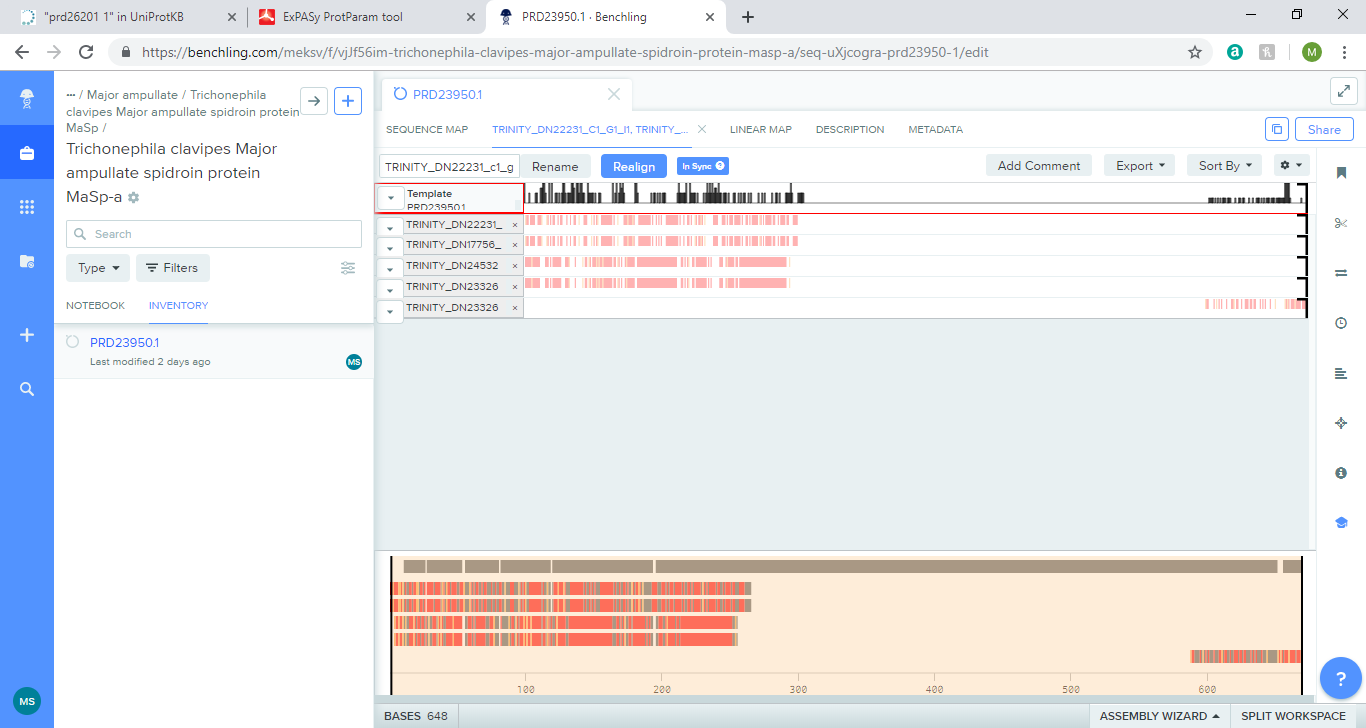
All of the sequences seem to fall in the N-terminus. Most of the percentages of identical matches is less than 50% than and the number of mismatches compared to alignment length suggests that the sequence alignment is very poor. Only TRINITY\_DN22395\_c1\_g1\_i1 is slightly better than most of the other query sequences with a percentage match over 50%.

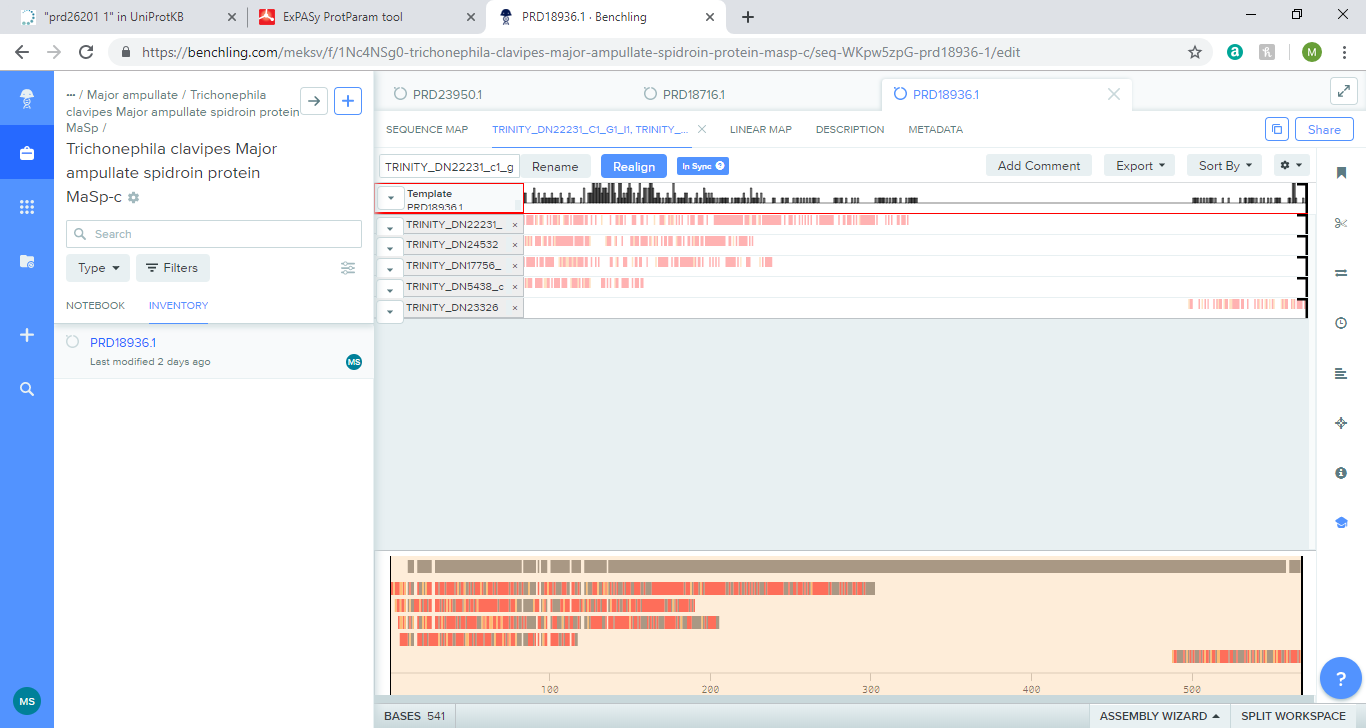
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | TuSp, N-term | 24532\_c2\_g1\_i1, 28-150bp | 22231\_c1\_g1\_i1, 28-150bp | 22395\_c0\_g1\_i1, 28-150bp | 5438\_c1\_g1\_i1 | 17756\_c0\_g2\_i2, 0-170bp |
| Ala | 8.9 | 8.9 | 9.7 | 14.6 | 9 | 8.8 |
| Gly | 5.7 | 4.1 | 8.6 | 8.9 | 2.8 | 7.6 |
| Ser | 20.3 | 13 | 16.1 | 10.6 | 12.5 | 14.1 |
| Leu | 12.2 | 4.9 | 3.2 | 8.1 | 8.3 | 5.9 |
| Lys | 0.8 | 6.5 | 1.1 | 0.8 | 4.2 | 1.8 |
| Met | 0 | 4.1 | 6.5 | 0.8 | 2.8 | 7.6 |

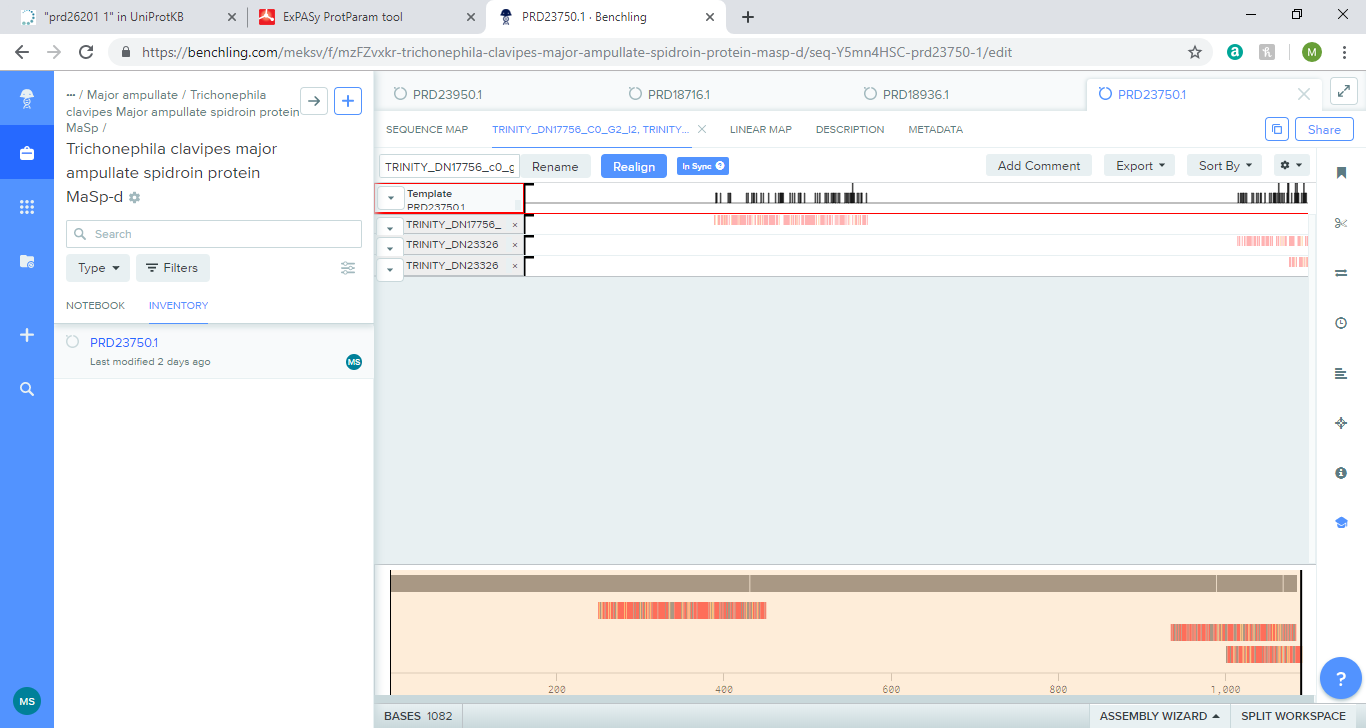
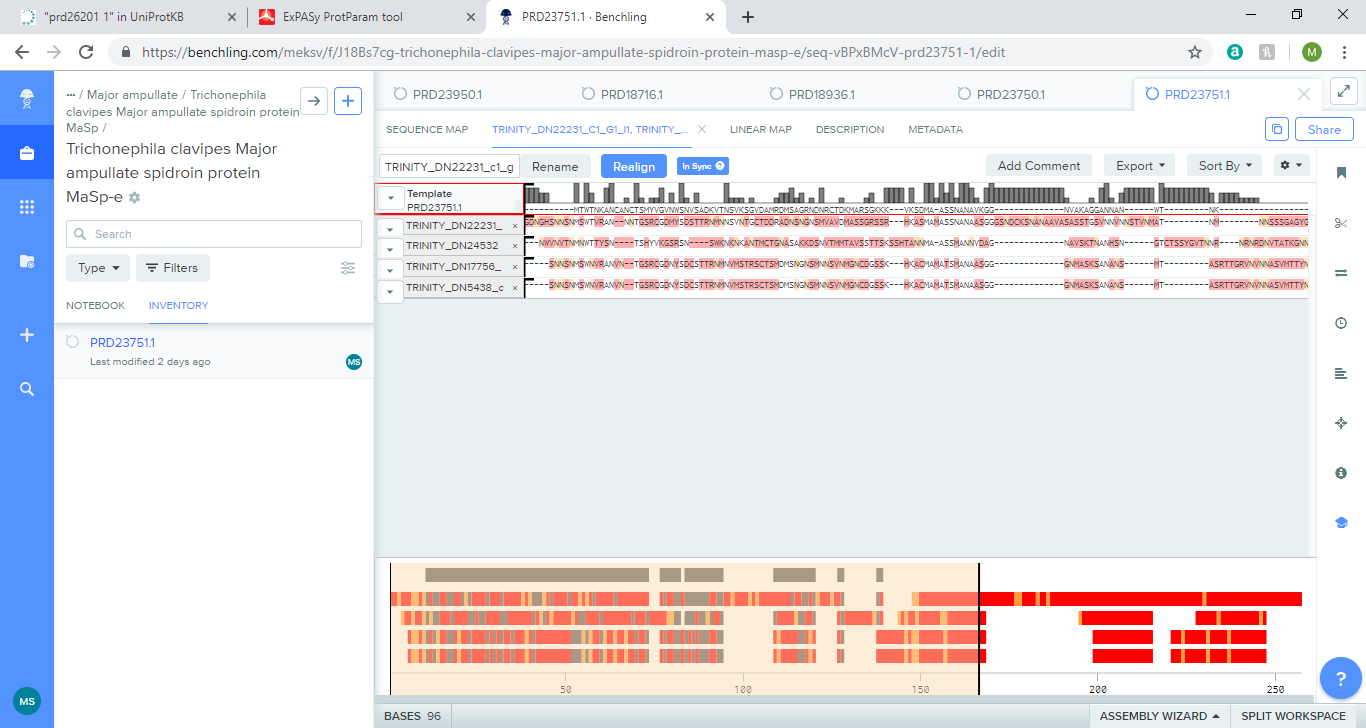
**Table 18:** Amino acid composition (%) of Trichonephila clavipes Tubuliform spidroin protein, TuSp. Yellow box means equal or more than 5%. Orange box means equal or more than 10%.

Table 18 shows certain amino acid compositions of the template sequences and the aligned sequences where there is a slight difference. Ser is very high in most sequences; Ala is high in most sequences except it is slightly higher in the 28-150bp region of TRINTY\_DN22395\_c0\_g1\_i1. Gly has a lower percentage than both of those amino acids, but it still high. Leu has quite a low percentage in two of the query sequences, TRINITY\_DN24532\_c2\_g1\_i1 and TRINITY\_DN22231\_c1\_g1\_i1 (more than 5% difference) compared to the template sequence. Lys is higher in the 28-150bp query sequence of TRNITY\_DN24532\_c2\_g1\_i1 (more than 5%) than the N-term of the template sequence. Met is higher in the 28-150bp region of TRNITY\_DN22231\_c1\_g1\_i1 and 0-170bp region of TRNITY\_DN17756\_c0\_g2\_i2 than the N-term of the template sequence (more than 5% difference).

**3.3.5 Major ampullate**

**3.3.5.1 Trichonephila clavipes Major ampullate spidroin protein MaSp**





**Figure 21:** Sequence alignment of Trichonephila clavipes Major ampullate spidroin protein MaSp. From top to bottom: MaSp-a, MaSp-b, MaSp-c, MaSp-d, MaSp-e. Red lines mean different nucleotide. Spaces between red lines means same nucleotide.

+

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Query ID** | **E value** | **Bit score** | **Percentage of identical matches** | **Alignment length** | **Number of mismatches** | **Number of gap openings** |
| **MaSp-a** |  |  |  |  |  |  |
| TRINITY\_DN22231\_c1\_g1\_i1 | 2.84E-31 | 117 | 41.406 | 128 | 75 | 0 |
| TRINITY\_DN17756\_c0\_g2\_i2 | 5.91E-30 | 111 | 37.857 | 140 | 87 | 0 |
| TRINITY\_DN24532\_c2\_g1\_i1 | 1.39E-25 | 99.4 | 38.037 | 163 | 92 | 5 |
| TRINITY\_DN23326\_c1\_g1\_i1 | 5.64E-21 | 84 | 53.409 | 88 | 41 | 0 |
| TRINITY\_DN23326\_c1\_g2\_i1 | 5.68E-18 | 72.4 | 48.352 | 91 | 47 | 0 |
| **MaSp-b** |  |  |  |  |  |  |
| TRINITY\_DN22231\_c1\_g1\_i1 | 3.11E-35 | 125 | 46.104 | 154 | 81 | 2 |
| TRINITY\_DN24532\_c2\_g1\_i1 | 2.16E-19 | 80.9 | 33.516 | 182 | 109 | 5 |
| TRINITY\_DN17756\_c0\_g2\_i2 | 9.77E-35 | 122 | 43.506 | 154 | 85 | 2 |
| TRINITY\_DN5438\_c1\_g1\_i1 | 7.60E-17 | 70.9 | 31.884 | 138 | 84 | 4 |
| TRINITY\_DN23326\_c1\_g2\_i1 | 2.75E-15 | 64.7 | 46.154 | 78 | 42 | 0 |
| **MaSp-c** |  |  |  |  |  |  |
| TRINITY\_DN22231\_c1\_g1\_i1 | 2.71E-34 | 125 | 46.104 | 154 | 81 | 2 |
| TRINITY\_DN24532\_c2\_g1\_i1 | 1.57E-21 | 87.4 | 32.335 | 167 | 101 | 5 |
| TRINITY\_DN17756\_c0\_g2\_i2 | 6.10E-34 | 122 | 43.506 | 154 | 85 | 2 |
| TRINITY\_DN5438\_c1\_g1\_i1 | 1.09E-16 | 70.5 | 31.884 | 138 | 84 | 4 |
| TRINITY\_DN23326\_c1\_g2\_i1 | 6.29E-15 | 66.2 | 50.667 | 75 | 37 | 0 |
| **MaSp-d** |  |  |  |  |  |  |
| TRINITY\_DN23326\_c1\_g1\_i1 | 1.06E-17 | 74.3 | 53.012 | 83 | 39 | 0 |
| TRINITY\_DN23326\_c1\_g2\_i1 | 1.39E-12 | 57 | 41.86 | 86 | 50 | 0 |
| **MaSp-e** |  |  |  |  |  |  |
| TRINITY\_DN22231\_c1\_g1\_i1 | 1.78E-13 | 61.2 | 33.083 | 133 | 85 | 3 |
| TRINITY\_DN24532\_c2\_g1\_i1 | 1.42E-06 | 41.2 | 26.154 | 130 | 86 | 4 |
| TRINITY\_DN17756\_c0\_g2\_i2 | 7.84E-12 | 55.8 | 30.075 | 133 | 89 | 3 |
| TRINITY\_DN5438\_c1\_g1\_i1 | 1.73E-08 | 45.4 | 26.852 | 108 | 71 | 3 |

**Table 19:** Blastp of Trichonephila clavipes Major ampullate spidroin protein MaSp.The table shows data for the first and last query sequence that aligns with the template sequence.

Given that the template sequences are isoforms, some of the aligned sequences are the same. Some of the sequences (TRINITY\_DN23326\_c1\_g1\_i1 and TRINITY\_DN23326\_c1\_g2\_i1) have a percentage of identical match above 50%, which is distinct and good for a protein sequence. Most of the sequences seem to fall in the N-term region while some fall in the C terminus.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | MaSp-a, N-term (28-152bp) | MaSp-b, N-term (28-152bp) | MaSp-c, N-term (28-152bp) | MaSp-e, N-term (28-152bp) | 22231\_c1g1i1 (28-152bp) | 17756\_c0g2i2 (28-152bp) | 24532\_c2g1i1 (28-152bp) | 23326\_c1g1i1 (28-152bp) | 5438\_c1g1i1 (28-152bp) |
| Ala | 12 | 17.6 | 17.6 | 11.7 | 11.2 | 8.8 | 8.8 | 5.6 | 11.1 |
| Gly | 4.8 | 4.8 | 4.8 | 6.8 | 7.2 | 6.4 | 4 | 12 | 3.4 |
| Ser | 16 | 12 | 12 | 8.7 | 16.8 | 13.6 | 12.8 | 25.6 | 12 |
| Gln | 4 | 5.6 | 5.6 | 6.8 | 8.8 | 9.6 | 5.6 | 1.6 | 6 |
| Met | 5.6 | 4.8 | 4.8 | 3.9 | 4.8 | 9.6 | 4 | 0.8 | 2.6 |
| Val | 4 | 4 | 4 | 6.8 | 5.6 | 4 | 4.8 | 11.2 | 6.8 |

**Table 20:** Amino acid composition (%) of the N-term Trichonephila clavipes Major ampullate spidroin protein MaSp. Yellow box means equal or more than 5%. Orange box means equal or more than 10%.

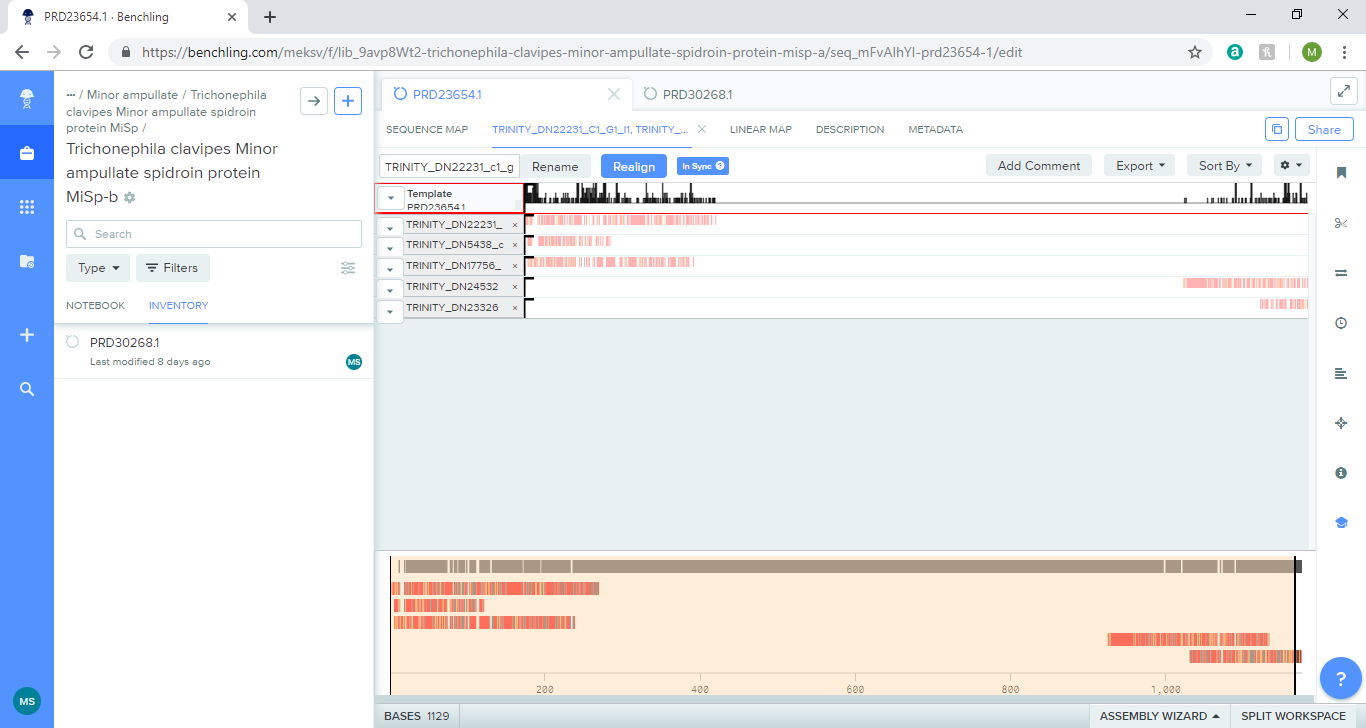
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MaSp-a (570-640bp) | MaSp-b (250-325bp) | MaSp-c (470-540bp) | MaSp-d (940-1080bp) | 23326\_c1g1i1 | 23326\_c1g2i1 |
| Ala | 12.7 | 27.6 | 31 | 17 | 8.9 | 7.9 |
| Gly | 31 | 34.2 | 46.5 | 41.1 | 17.8 | 4 |
| Ser | 12.7 | 13.2 | 1.4 | 12.8 | 25.4 | 25.7 |
| Val | 5.6 | 6.6 | 2.8 | 1.4 | 9.5 | 9.9 |
| Gln | 9.9 | 6.6 | 8.5 | 1.4 | 1.8 | 3 |
| Ile | 1.4 | 0 | 0 | 0 | 5.3 | 7.9 |
| Leu | 0 | 3.9 | 4.2 | 0 | 8.3 | 9.9 |

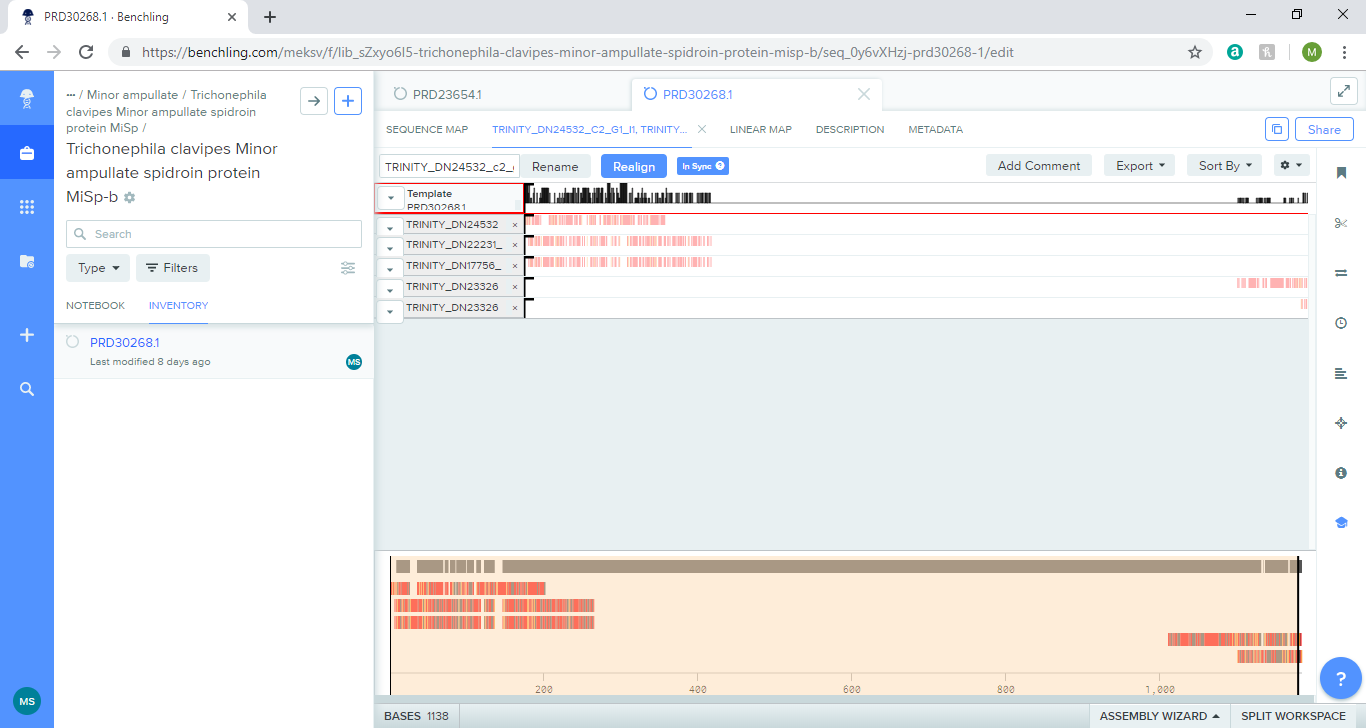
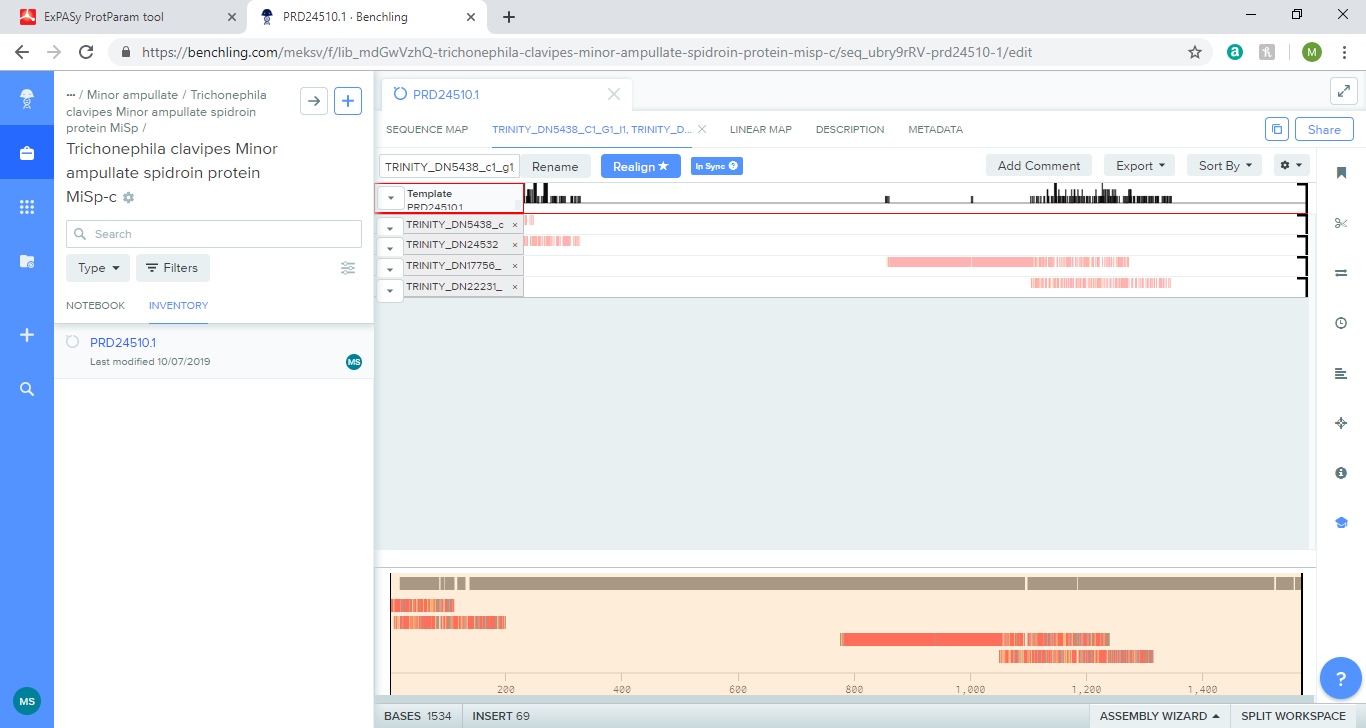
**Table 21:** Amino acid composition (%) of the C-term Trichonephila clavipes Major ampullate spidroin protein MaSp. Yellow box means equal or more than 5%. Orange box means equal or more than 10%.

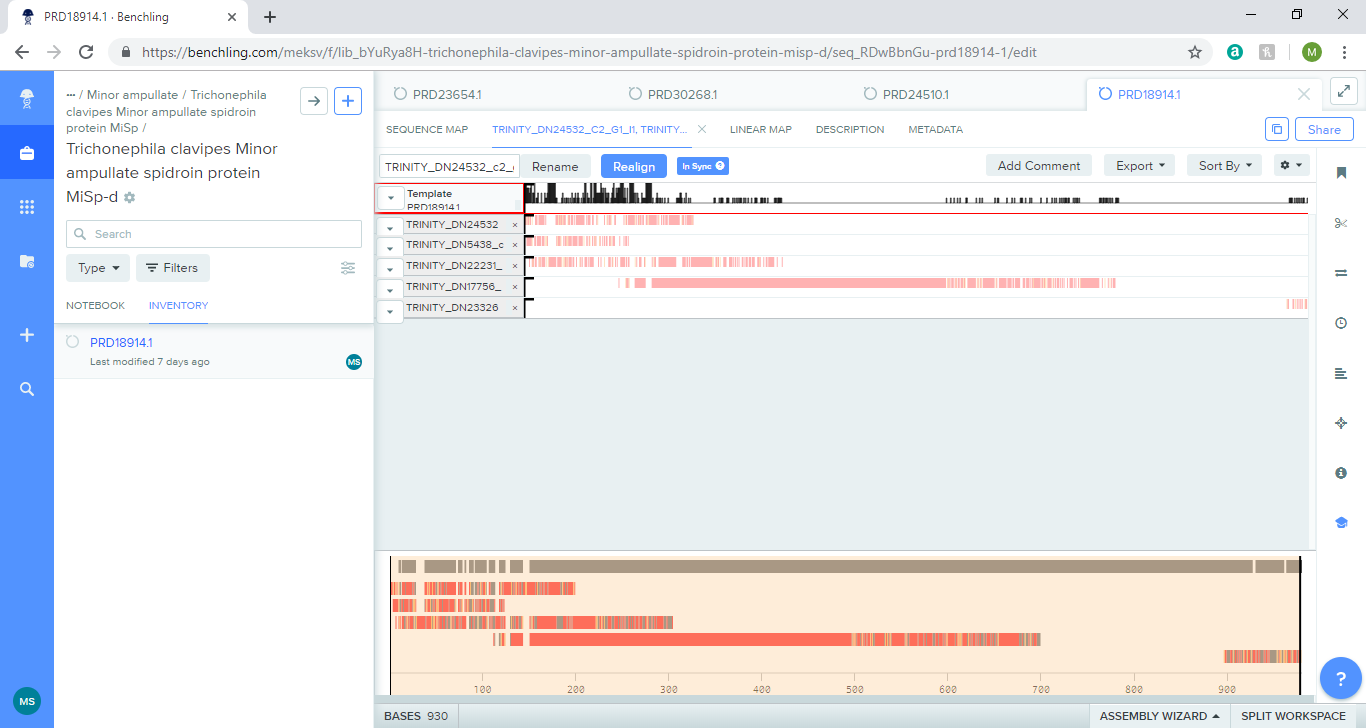
Table 20 and Table 21 shows certain amino acid compositions of the template sequences and the aligned sequences where there is a slight difference for MaSp N term region and possible C -term. For the N-term, the percentage of Ala in the template sequences is higher compared to all the query sequences (above 10%), except TRINITY\_DN5438\_c1\_g1\_i1. The percentage of Gly is higher in the query sequence TRINTY\_DN23326\_c1\_g1\_i11 than the template sequences and other query sequences. Ser is very high in most sequences (above 10%) except for the N-term of MaSp-e. Met and Gln is lower in the query sequence TRINITY\_DN23326\_c1\_g1\_i1 than the other sequences, resulting in a range of 0.8-9.6% and 1.6-9.6%, respectively. Val is higher in TRNITY\_DN23326\_c1\_g1\_i1 than other template sequences and other query sequences. For the possible C -term, Ala and Gly is very high in the template sequences but not as high in query sequences. Ser is much higher in the query sequences than the template sequence. The other amino acids listed show that there is range of percentages even between the template sequences. It is unlikely that this is the C-term region.

**3.3.6 Minor ampullate**

**3.3.6.1 Trichonephila clavipes Minor ampullate spidroin protein, MiSp**







**Figure 22**: Sequence alignment of Trichonephila clavipes Major ampullate spidroin protein MaSp. From top to bottom: MaSp-a, MaSp-b, MaSp-c, MaSp-d, MaSp-e. Red lines mean different nucleotide. Spaces between red lines means same nucleotide.

+

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Query ID** | **E value** | **Bit score** | **Percentage of identical matches** | **Alignment length** | **Number of mismatches** | **Number of gap openings** |
| **MiSp-a** |  |  |  |  |  |  |
| TRINITY\_DN22231\_c1\_g1\_i1 | 7.19E-28 | 107 | 38.889 | 144 | 84 | 2 |
| TRINITY\_DN5438\_c1\_g1\_i1 | 4.88E-15 | 65.9 | 29.927 | 137 | 82 | 3 |
| TRINITY\_DN17756\_c0\_g2\_i2 | 9.32E-27 | 102 | 36.806 | 144 | 87 | 1 |
| TRINITY\_DN24532\_c2\_g1\_i1 | 5.29E-20 | 83.2 | 32 | 150 | 92 | 4 |
| TRINITY\_DN23326\_c1\_g1\_i1 | 2.78E-14 | 64.7 | 51.961 | 102 | 49 | 0 |
| **MiSp-b** |  |  |  |  |  |  |
| TRINITY\_DN24532\_c2\_g1\_i1 | 5.65E-19 | 80.1 | 31.579 | 152 | 94 | 3 |
| TRINITY\_DN22231\_c1\_g1\_i1 | 3.52E-29 | 111 | 39.073 | 151 | 88 | 2 |
| TRINITY\_DN17756\_c0\_g2\_i2 | 6.72E-26 | 100 | 35.762 | 151 | 93 | 1 |
| TRINITY\_DN23326\_c1\_g1\_i1 | 1.30E-13 | 62.8 | 46.528 | 144 | 66 | 2 |
| TRINITY\_DN23326\_c1\_g2\_i1 | 4.13E-15 | 64.3 | 57.333 | 75 | 32 | 0 |
| **MiSp-c** |  |  |  |  |  |  |
| TRINITY\_DN5438\_c1\_g1\_i1 | 1.33E-13 | 61.6 | 34.615 | 104 | 61 | 2 |
| TRINITY\_DN24532\_c2\_g1\_i1 | 7.12E-19 | 80.1 | 32.667 | 150 | 91 | 3 |
| **MiSp-d** |  |  |  |  |  |  |
| TRINITY\_DN24532\_c2\_g1\_i1 | 1.30E-17 | 76.3 | 32.667 | 150 | 91 | 3 |
| TRINITY\_DN5438\_c1\_g1\_i1 | 4.65E-14 | 63.2 | 32.609 | 138 | 77 | 4 |
| TRINITY\_DN22231\_c1\_g1\_i1 | 5.24E-23 | 93.6 | 38.194 | 144 | 85 | 2 |
| TRINITY\_DN17756\_c0\_g2\_i2 | 2.03E-23 | 92.8 | 37.5 | 144 | 86 | 1 |
| TRINITY\_DN23326\_c1\_g2\_i1 | 9.83E-13 | 57.4 | 51.579 | 95 | 46 | 0 |

**Table 22:** Blastp of Trichonephila clavipes Major ampullate spidroin protein MaSp.The table shows data for the first and last query sequence that aligns with the template sequence.

Given that the template sequences are isoforms, some of the aligned sequences are the same. Some of the sequences (TRINITY\_DN23326\_c1\_g1\_i1 and TRINITY\_DN23326\_c1\_g2\_i1) have a percentage of identical match above 50%, which is distinct and good for a protein sequence. Most of the sequences seem to fall in the N-term region while some fall in the C terminus.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | MiSp-a, N-term, 26-146bp | MiSp-b, N-term, 26-144bp | MiSp-c, N-term, 26-146bp | MiSp-d, N-term, 26-146bp | 22231\_c1g1i1 | 5438\_c1g1i1 | 17756\_c0g2i2 | 24532\_c2g1i1 |
| Ala | 14.9 | 11.8 | 11.6 | 12.4 | 11.6 | 9 | 9.2 | 9.2 |
| Gly | 5 | 9.2 | 9.1 | 9.1 | 7.4 | 2.8 | 6.7 | 4.2 |
| Ser | 12.4 | 8.4 | 9.1 | 9.9 | 16.5 | 12.5 | 14.2 | 13.3 |

**Table 23:** Amino acid composition (%) of the N-term Trichonephila clavipes Minor ampullate spidroin protein MiSp. Yellow box means equal or more than 5%. Orange box means equal or more than 10%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MiSp-a, 920-1260bp | MiSp-b, 970-1283bp | MiSp-d, 850-1949bp | 24532\_c2g1i1 | 23326\_c1g1i1 | 23326\_c1g2i1 |
| Ala | 14.7 | 19.4 | 22.5 | 11.1 | 8.9 | 7.9 |
| Gly | 15 | 25.2 | 27.5 | 4.4 | 17.8 | 4 |
| Ser | 19.6 | 16.2 | 13.5 | 15 | 25.4 | 25.7 |
| Asn | 5 | 1.6 | 0.5 | 4.9 | 7.7 | 9.9 |
| Glu | 1.8 | 1.3 | 1.5 | 6.6 | 1.8 | 2 |
| Thr | 6.5 | 6.1 | 2.5 | 11.5 | 2.4 | 2 |

**Table 24:** Amino acid composition (%) of the C-term Trichonephila clavipes Minor ampullate spidroin protein MiSp. Yellow box means equal or more than 5%. Orange box means equal or more than 10%.

Table 22 and 23 shows certain amino acid compositions of the template sequences and the aligned sequences where there is a slight difference. For the N-term region, Ser is very high in most sequences but in this protein type, the amino acid composition of Ala is higher, which is fairly unique. Gly has a lower percentage than both of those amino acids, but it still high. For other amino acids, there are some differences but there is no distinct change (more than 5%). For the possible C-term, Ala is higher in the template sequences than the query sequences. For Gly, except TRINITY\_DN23326\_c1\_g1\_i1 , there is a higher percentage in the template sequence than the query sequences. Ser is very high in the template and query sequences. There is a wide range for Asn in the sequences with MiSp-d having 0.5% of Asn, while TRINITY\_DN23326\_c1\_g2\_i1 is 9.9% of Asn. Glu is slightly higher in TRINITY\_DN24532\_c2\_g1\_i1 than the template sequences and the other query sequences. Thr is higher in TRINITY\_DN24532\_c2\_g1\_i1 than the template sequence than the other query sequences.

**Discussion**

Based on the origin of the template protein and nucleotide sequences found using NCBI and BLAST+, there appears to be an evolutionary relationship between Eratigena atrica and Araneus ventricous, Argiope argentata, Argiope bruennichi, Latrodectus Hesperus, Trichonoephila clavipes, at least when it comes to the production of silk. There is quite a lot of sequence alignment between Eratigena atrica and Latrodectus Hesperus, most likely due to a lot of research being done on that species. The N-terminus and C-terminus of silk proteins should be conserved between species and while the sequence alignment does not match, the amino acid composition shows that it is fairly similar. This suggests that while the amino acids are not exactly in the same position of the protein sequence, the abundance of amino acids being conserved matches with what is known with Ser being quite high with Ala and Gly being fairly high as well. In some cases, the percentage of Ala was more than Ser such as the major ampullate. The table below outlines what assembled transcripts match with the core fibre proteins and should be looked at in more detail. Also, most of the alignment was for the N-term rather than C term but could be since C term not being specified as a search term for NCBI. The table below outlines the type of silk and the corresponding template sequence to be focused on for future research. Some of the sequences are in more than one silk type.

|  |  |  |
| --- | --- | --- |
| **Silk** | N-term | C-term |
| Major ampullate | 22231\_c1g1i1, 17756\_c02g2i2, 24532\_c2g1i1, 23326\_c1g1i1,  5438\_c1g1i1 | 23326\_c1g1i1, 23326\_c1g2i1 |
| Minor ampullate | 22231\_c1g1i1, 5438\_c1g1i1,  17756\_c0g2i2, 24532\_c2g1i1 | 24532\_c1g1i1, 23326\_c1g2i1, 24532\_c2g1i1 |
| Flagelliform | 22395\_c0g1i1, 23891\_c1g3i1, 42663\_c0g1i1, 17756\_c0g2i2,  5438\_c1g1i1, 22231\_c1g1i1 | - |
| Aciniform | 24532\_c2g1i1,  5438\_c1g1i1, 22231\_c1g1i1 | 13506\_c0\_g1\_i1 |
| Tubuliform | 24532\_c2g1i1, 22231\_c1g1i1,  22395\_c0g1i1,5438\_c1g1i1,  17756\_c0g2i2 | -- |
| Aggregate | 22231\_c1g1i1, 5438\_c1g1i1, 24532\_c2g1i1 | - |
| Egg case | 22395\_c0g1i1 | 24532\_c2g1i1 |

**Table 25:** Assembled transcripts from our data compared to the silk type

**References**

Rising, A., Nimmervol, H., Fernadez-Arias, A., Storckenfeldt, E., Knight, D.P., Vollrath, F., Engström (2005). Spider Silk Proteins – Mechanical Property and Gene Sequence. *Zoological Science,* 22(3), pp.273-281

Hu, X., Vasanthavada, K., Kohler, K., McNary, S., Moore, A.M.F. and Vierra, C.A. (2006). Molecular mechanisms of spider silk. *Cellular and Molecular Life Sciences*, 63(17), pp.1986–1999.

BLAST® Command Line Applications User Manual [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2008