M1 GENIOMHE 2024/25: Project

Structural Genomics of *Triticum aestivum*

Group members: Aya BEN TAGHALINE, Joelle ASSY, Rayane ADAM

Table of contents

Introduction

Triticum aevistum (commonly known as wheat), is a complex eukaryotic organism belonging to kingdom Plantae, phylum Angiosperms, class Monocots, order Poales, family Poaceae, genus Triticum. This plant has been considered as one of the most important crops in the world, providing a staple food source for billions of people as it is mainly used to make bread.

Even though it is somehow considered a model organism in plant biology, it has a complex genome structure that makes it difficult to study. It is a hexaploid species with a large genome, consisting of 7n chromosomes and a high rate of repeats and transposable elements. This polyploidy is in fact derived from the hybridization of three different species due to an evolutionary event that occurred around 8,500–9,000 years ago. It comes from a tetraploid species having BBAA chromosomes and a diploid species having DD chromosomes. The tetraploid species is believed to be a free-threshing species and can be thought to be *Triticum monococcum* or *Triticum durum*, whereas the diploid species is *Aegilops tauschii*. Ending up with a hexaploid species with BBAADD chromosomes.[1]

[1] Levy, Avraham A., and Moshe Feldman. "Evolution and origin of bread wheat." The

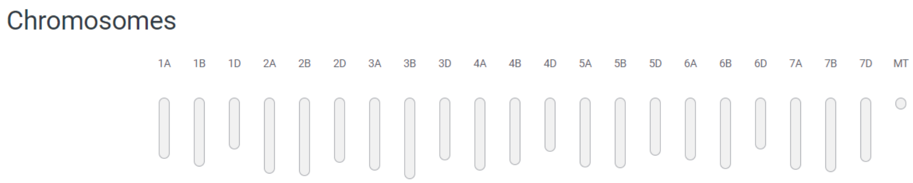


Figure 1: T. aestivum set of chromosomes from RefSeq

The goal of this project is to annotate a specific region of the genome of *Triticum aestivum* (wheat), mainly structurally annotate, using bioinformatics tools. The region of interest is a 14,001 bp sequence (`region8`), which we will analyze to predict genes, transposable elements, and other features. This task is considerably a hard one taking into consideration this complicated genome structure from polyploidy and the richness of repetitive elements, as well as its large size.
In this project, we will start of by a minor exploration fo our region then we'll perform gene prediction using a variety of tools then analyze and validate these results. We will also look for transposable elements in the region and perform a final annotation of the region to conclude with this report. We have used online servers, databases, api calls, unix tools, visualization software, python & bash scripting to perform the analysis. Supplementary results, data, code, figures and documentation can be found on the github repository of this project: github.com/raysas/wheat-seq-annotation.

Tools, databases and utils used in this project are listed in the following table by alphabetical order:

Name	Type	Use
AUGUSTUS	Webserver	Gene prediction tool
Artemis	Software	Visualizing genome annotation
Biopython	Python library	Retrieve data online via API calls, process sequences, and shift formats
Blast+	Unix tool	Blast locally against built databases from Uniprot proteome data
BWA-MEM2	Galaxy server	Map genomic regions against the reference genome
BEDTools	Unix tool	Manipulate BED files
Censor	Webserver	Annotate TEs
DNASubway	Webserver	Annotation pipeline to verify results
ENA database	Database	Retrieve TSA information and study data
Genome Data Viewer	website	Visualize mapped regions on chromosomes

Plant Cell 34.7 (2022): 2549-2567.

Name	Type	Use
ENSEMBL Plants	Database	Reference sequence and cDNA (transcripts) of the wheat genome
FastqGroomer	Galaxy	Standardize FASTQ format for mapping
FastQC	Galaxy	Check transcriptome quality
FGENESH	Webserver	Gene prediction tool
Galaxy EU	Webserver	Perform large-scale genomic analysis
GENEID	Webserver	Gene prediction tool
IGB	Software	Genome visualization
RefSeq	Database	Reference sequence of wheat genome chromosomes
RepeatMasker	Unix tool	Annotate TEs against the built Trep database
RNAStar	Galaxy	Splice-aware RNA-seq mapper
SAMtools	Unix tool	Manipulate SAM and BAM files
Trep	Database	TE database to run RepeatMasker locally
NCBI SRA	Database	Download RNA-seq raw fastq data
NCBI Nucleotide	Database	Sequence database
Uniprot	Database	Retrieve species proteome
FastQC	Galaxy	Check transcriptome quality
FastqGroomer	Galaxy	Standardize FASTQ format
IGB	Software	Genome visualization
RNAStar	Galaxy	Splice-aware RNA-seq mapper

Project met:

- ☒ annotate genes with complete coordinates, validation by the presence of transcribedsequences and/or homologous genes
- ☒ annotate proteins, potential protein functions, motifs and domains
- ☒ annotate transposable elements coordinate and family

and additionally:
- [x] localized the region on the reference genome, chromosome number and strand

Exploration

Sequence properties

Checking GC content in this region to have an idea about potential gene desnitites. For that we run the script:

```
$ python src/GCcontent.py data/region8.fasta
0.48
```

The GC content of the DNA sequence is 48%.

We proceed to see the length of the sequence:

```
$ expr $(tail -n +2 data/region8.fasta | wc -c) - $(tail -n +2 data/region8.fasta | wc -l)
14001
```

region8 is 14,001 bases long.

Region localization

Want to localize this region by mapping agaisnt the reference sequence of *Triticum aestivum* (available on RefSeq at GCF_018294505.1), which consists of 7*n* chromosomes. After retrieving the reference sequence, we perfomed mapping through Burrows-Wheeler Aligner MEM (bwa-mem) algorithm, and due to large genome size, we did this step on Galaxy because of the large computation time and memory required.

```
$ bwa-mem2 mem -t 4 data/sequences/reference/GCF_018294505.1_genomic.fna \
    data/region8.fasta > data/sequences/alignment/region8.sam
$ samtools sort data/sequences/alignment/region8.sam \
    > data/sequences/alignment/region8_aln.bam
$ bedtools bamtobed -i data/sequences/alignment/region8_aln.bam \
    > data/sequences/alignment/region8_aln.bed
```

We chose this mapper because it's perfect for medium length reads ranging between 100bp and megabases, in our case it's a 14kb sequence, keeping default parameters

Now we have in the output a `.bam` file and a `.bed` file. From the `.bam` file we can get the following information when running the following command:

```
$ samtools view -c -F 4 data/sequences/alignment/region8_aln.bam
```

```
region8 16      NC_057805.1     497158671       60      9565M1I4435M    *       0       0
<sequence>   *  NM:i:5   MD:Z:9565A1651C131G821T1828     AS:i:13973      XS:i:2788
```

From the `.bam` output we can see[2]:

- The CIGAR string `9565M1I4435M`, means that the read is 9565 bases long, then there is an insertion of 1 base, and then 4435 more bases.
- The `NM:i:5` field indicates that there are 5 mismatches in the alignment.
- The `MD:Z:9565A1651C131G821T1828` field indicates the mismatches in the alignment.

If we further proceed conversion onto a `.bed` file, we get the following info:

```
NC_057805.1 497158670   497172670   region8 60  -
```

This means that the region8 is:

- located on the chromosome `NC_057805.1`
- position starting from `497158670` and ending at `497172670`

[2] Li, Heng, et al. "The sequence alignment/map format and SAMtools." bioinformatics 25.16 (2009): 2078-2079.

- on the negative strand.

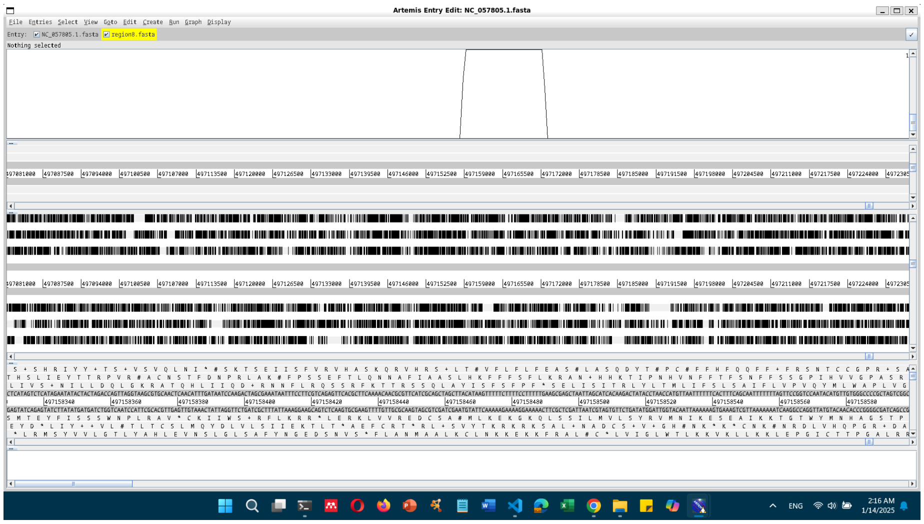

Artemis Entry Edit: NC_057805.1.fasta
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ NC_057805.1.fasta ☑ region8.fasta
Nothing selected
2:16 AM
1/14/2025

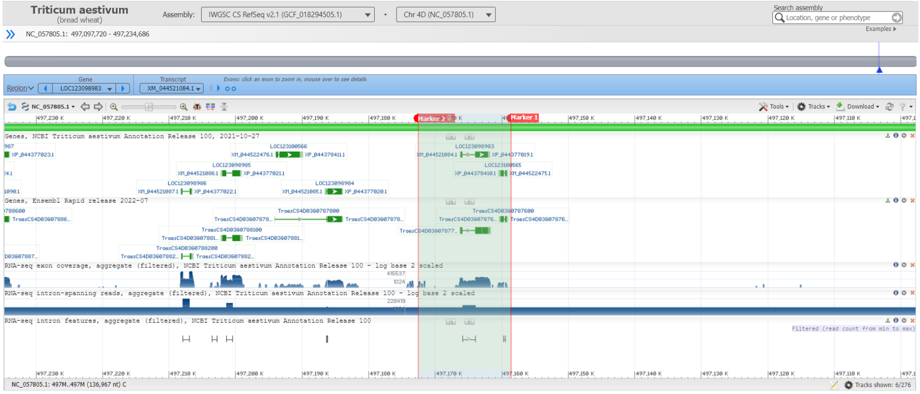
Figure 2: Alignment of region8 of chromosome 4D using bam output file

Reflection: our sequence is of length 14469, and the read is 9565+1+4435=14001, which means that the alignment is EXACTLY the same length as the sequence, and the 5 mismatches are not significant relative to the number of bases. We can thus infer that region8 is well mapped to the reference genome on the negative strand of chromosome `NC_057805.1` starting at position `497158670` and ending at `497172670`. And according to the table in [3] retrieved from RefSeq, this

[3]

Chromosome	GenBank	RefSeq	Size (bp)	GC content (%)	Unlocalized count	Action
1A	CM031178.1	NC_057794.1	598,660,471	46	0	
1B	CM031179.1	NC_057795.1	700,547,350	46	0	
1D	CM031180.1	NC_057796.1	498,638,509	46.5	0	
2A	CM031181.1	NC_057797.1	787,782,082	46	0	
2B	CM031182.1	NC_057798.1	812,755,788	46	0	
2D	CM031183.1	NC_057799.1	656,544,405	46.5	0	
3A	CM031184.1	NC_057800.1	754,128,162	46	0	
3B	CM031185.1	NC_057801.1	851,934,019	46	0	
3D	CM031186.1	NC_057802.1	619,618,552	46.5	0	
4A	CM031187.1	NC_057803.1	754,227,511	46	0	
4B	CM031188.1	NC_057804.1	673,810,255	46.5	0	
4D	CM031189.1	NC_057805.1	518,332,611	46.5	0	
5A	CM031190.1	NC_057806.1	713,360,525	46	0	
5B	CM031191.1	NC_057807.1	714,697,677	46	0	
5D	CM031192.1	NC_057808.1	569,951,140	46.5	0	
6A	CM031193.1	NC_057809.1	622,669,697	46	0	
6B	CM031194.1	NC_057810.1	731,188,232	46.5	0	
6D	CM031195.1	NC_057811.1	495,380,293	46.5	0	
7A	CM031196.1	NC_057812.1	744,491,536	46	0	

chromosome is the 4D chromosome of *Triticum aestivum*.



Triticum aestivum (bread wheat)

Assembly: IWGSC CS RefSeq v2.1 (GCF_018294505.1) · Chr 4D (NC_057805.1)

NC_057805.1: 497,097,720 - 497,234,686

Gene: LOC123098983 Transcript: XM_044521084.1

Exons: click an exon to zoom in, mouse over to see details

Genes, NCBI Triticum aestivum Annotation Release 100, 2021-10-27

Genes, Ensembl Rapid release 2022-07

RNA-seq exon coverage, aggregate (filtered), NCBI Triticum aestivum Annotation Release 100 - log base 2 scaled

RNA-seq intron-spanning reads, aggregate (filtered), NCBI Triticum aestivum Annotation Release 100 - log base 2 scaled

RNA-seq intron features, aggregate (filtered), NCBI Triticum aestivum Annotation Release 100

Filtered (read count from min to max)

NC_057805.1: 497M..497M (136,967 nt) C

Tracks shown: 6/276

Figure 3: Start and End positions of the alignment on the reference genome - RefSeq Genome Browser (marker 1: start position; marker 2: end position)

We can visualize the `.bed` file in Ensembl Plants, IWGSC assembly converter

Moreover, this region has a GC content of 48% (as reported earlier), which is ~2% higher than the average GC content of the whole 4D chromosome (46.5%). This might indicate a high gene density in this region, as genes are known to have a higher GC content than the rest of the genome.

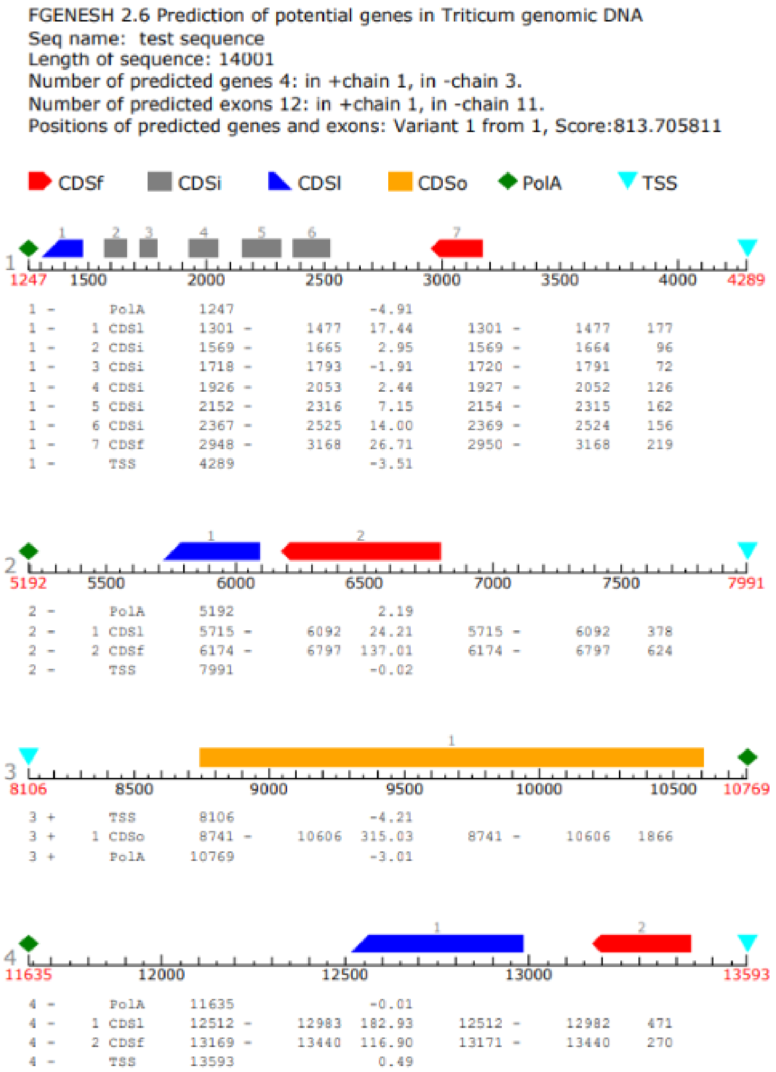
Gene Prediction

Tools

FGENESH

We used the FGENESH site, providing only the name of the organism, Triticum aestivum (wheat), and the DNA sequence. The default settings used a gene prediction model specifically trained for Triticum aestivum, which allowed the software to identify potential genes, exons, and other features such as transcription start sites (TSS) and polyadenylation sites (PolA). The output includes the positions of coding sequences : The parts of exons encoding proteins, TSS (Transcription Start Site): Where transcription begins, PolA (Polyadenylation Site): Where mRNA processing ends.
The FGENESH analysis of a 14,001 bp Triticum genomic DNA sequence predicted four genes, with one on the positive strand and three on the negative strand. In total, 12 exons were identified, with one on the positive strand and 11 on the

7B	CM031197.1	NC_057813.1	764,072,961	46	0
7D	CM031198.1	NC_057814.1	642,921,167	46.5	0
MT	EU534409.1	NC_036024.1	452,526	44.5	0



FGENESH 2.6 Prediction of potential genes in Triticum genomic DNA
Seq name: test sequence
Length of sequence: 14001
Number of predicted genes 4: in +chain 1, in -chain 3.
Number of predicted exons 12: in +chain 1, in -chain 11.
Positions of predicted genes and exons: Variant 1 from 1, Score:813.705811

```
 1 -       PolA     1247             -4.91
 1 -     1 CDSl     1301 -     1477  17.44     1301 -     1477    177
 1 -     2 CDSi     1569 -     1665   2.95     1569 -     1664     96
 1 -     3 CDSi     1718 -     1793  -1.91     1720 -     1791     72
 1 -     4 CDSi     1926 -     2053   2.44     1927 -     2052    126
 1 -     5 CDSi     2152 -     2316   7.15     2154 -     2315    162
 1 -     6 CDSi     2367 -     2525  14.00     2369 -     2524    156
 1 -     7 CDSf     2948 -     3168  26.71     2950 -     3168    219
 1 -       TSS      4289             -3.51

 2 -       PolA     5192              2.19
 2 -     1 CDSl     5715 -     6092  24.21     5715 -     6092    378
 2 -     2 CDSf     6174 -     6797 137.01     6174 -     6797    624
 2 -       TSS      7991             -0.02

 3 +       TSS      8106             -4.21
 3 +     1 CDSo     8741 -    10606 315.03     8741 -    10606   1866
 3 +       PolA    10769             -3.01

 4 -       PolA    11635             -0.01
 4 -     1 CDSl    12512 -    12983 182.93    12512 -    12982    471
 4 -     2 CDSf    13169 -    13440 116.90    13171 -    13440    270
 4 -       TSS     13593              0.49
```

Figure 4: FGENESH results

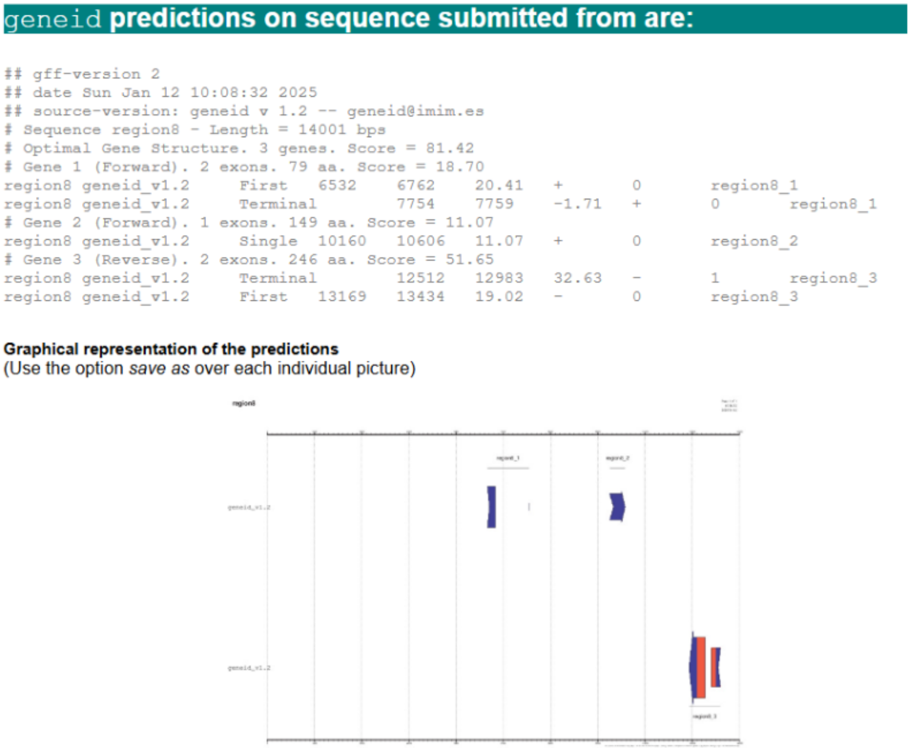
negative strand.
Gene 1, located on the negative strand, extends from position 1301 to 3168 and contains 7 exons, starting with the first coding exon (CDSl) at 1301–1477 and ending with the last exon (CDSf) at 2948–3168. It also includes 5 intermediate coding exons at positions 1569–1665, 1718–1783, 1926–2053, 2152–2316, and 2367–2525. The transcription start site (TSS) is identified at position 4289, while the polyadenylation site (PolA) is located at 1247. This gene translates into a protein of 340 amino acids.
Gene 2, also on the negative strand, extends from 5715 to 6797 and contains 2 exons, with the first coding exon (CDSl) at 5715–6092 and the last exon (CDSf) at 6174–6797, producing a protein of 333 amino acids. The Transcription start site (TSS) is identified at position 7991, while polyadenylation site (PolA) is located at 5192.
Gene 3, on the positive strand, is a single-exon gene (CDSo) located between 8741 and 10606, encoding a protein of 621 amino acids. The Transcription start site (TSS) is identified at position 8106, while polyadenylation site (PolA) is located at 10769.
Gene 4, on the negative strand, spans 12512–13440 with 2 exons; the first coding exon (CDSl) is at 12512–12983, and the final exon (CDSf) is at 13169–13440, translating into a protein of 247 amino acids. Transcription start site (TSS) is identified at position 13593, while polyadenylation site (PolA) is located at 11635.
The gene features, including exon positions and their strand orientation, suggest diverse transcriptional structures, with detailed sequences provided for both mRNA and proteins.

GENEID

Gene 1 is located on the forward strand (+) and consists of two exons, with the first exon positioned from 6532 to 6762 and the terminal exon from 7754 to 7759. Gene 2 is also on the forward strand (+) and is a single-exon gene, extending from 10160 to 10606. In contrast, Gene 3 is on the reverse strand (-) and has two exons, with the terminal exon located between 12512 and 12983, and the first exon from 13169 to 13434. The annotation reflects the strand orientation, with Gene 1 and Gene 2 being forward-strand genes, while Gene 3 is on the reverse strand, where exons are annotated in reverse order, starting from the terminal exon

AUGUSTUS

The AUGUSTUS gene prediction tool (version 3.3.3) analyzed a 14,001 bp sequence using the wheat parameter set and identified two genes, one on the forward strand and one on the reverse strand. Gene 1 on the forward strand, extends from 6226 to 10861 and contains two exons separated by an intron. The start codon is located in exon 1 (6532–6534), while the stop codon is in exon 2 (10604–10606). The coding sequence (CDS) includes two segments:

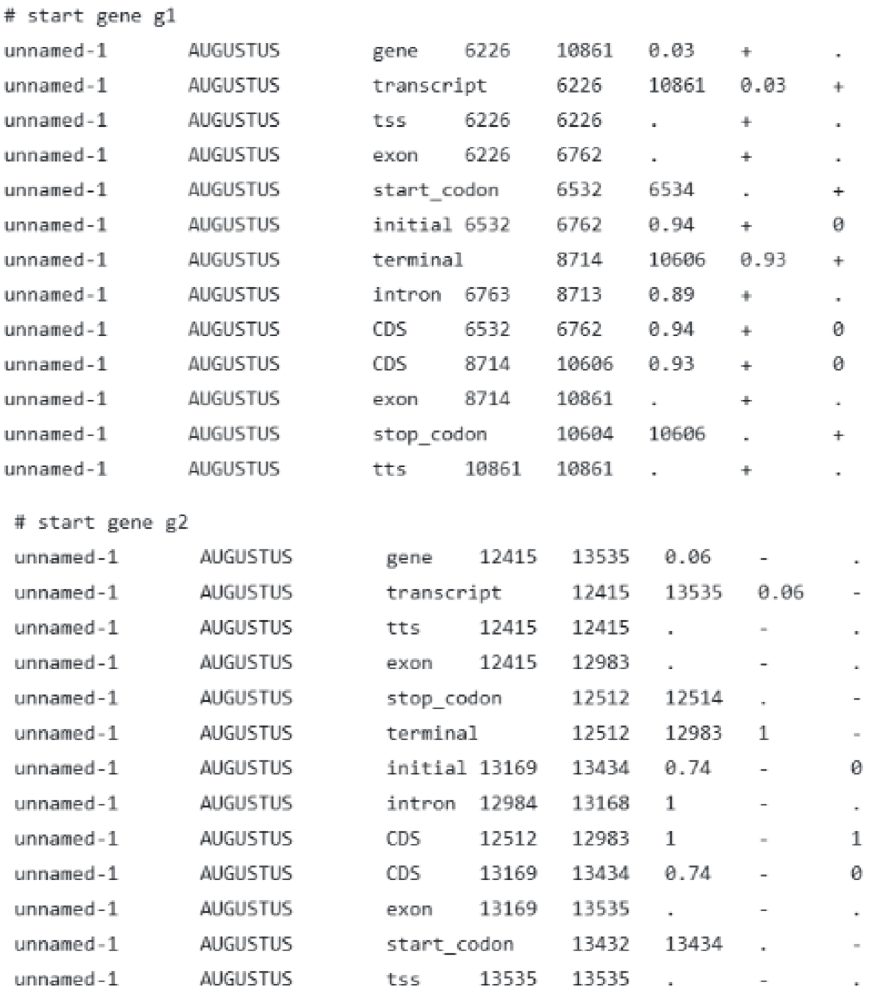


geneid predictions on sequence submitted from are:

```
## gff-version 2
## date Sun Jan 12 10:08:32 2025
## source-version: geneid v 1.2 -- geneid@imim.es
# Sequence region8 - Length = 14001 bps
# Optimal Gene Structure. 3 genes. Score = 81.42
# Gene 1 (Forward). 2 exons. 79 aa. Score = 18.70
region8 geneid_v1.2     First   6532    6762    20.41   +       0       region8_1
region8 geneid_v1.2     Terminal        7754    7759    -1.71   +       0       region8_1
# Gene 2 (Forward). 1 exons. 149 aa. Score = 11.07
region8 geneid_v1.2     Single  10160   10606   11.07   +       0       region8_2
# Gene 3 (Reverse). 2 exons. 246 aa. Score = 51.65
region8 geneid_v1.2     Terminal        12512   12983   32.63   -       1       region8_3
region8 geneid_v1.2     First   13169   13434   19.02   -       0       region8_3
```

Graphical representation of the predictions
(Use the option *save as* over each individual picture)

Figure 5: geneid results



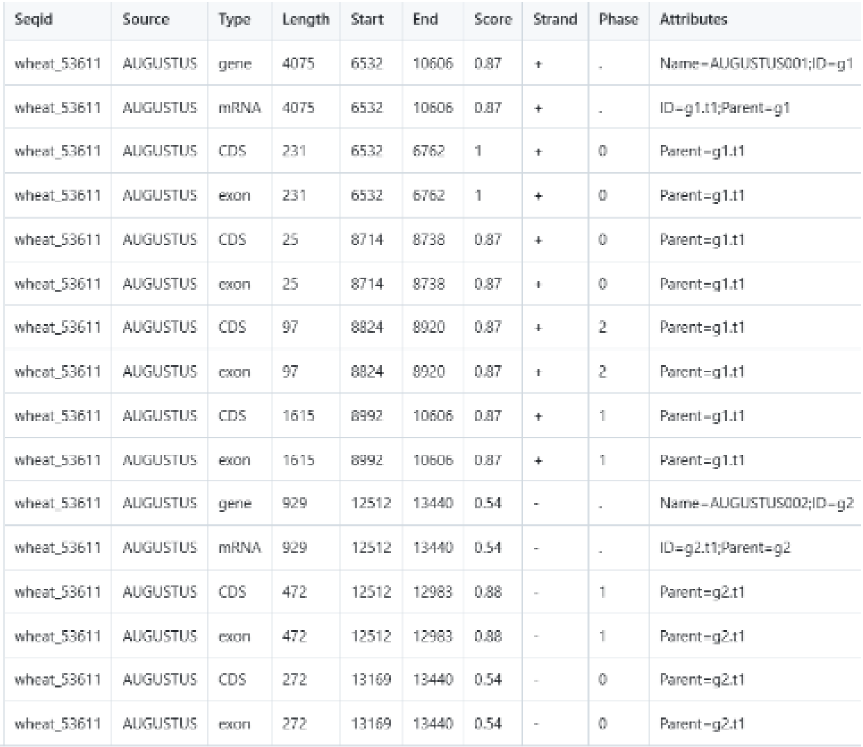
```
# start gene g1
unnamed-1       AUGUSTUS        gene    6226    10861   0.03    +       .
unnamed-1       AUGUSTUS        transcript      6226    10861   0.03    +
unnamed-1       AUGUSTUS        tss     6226    6226    .       +       .
unnamed-1       AUGUSTUS        exon    6226    6762    .       +       .
unnamed-1       AUGUSTUS        start_codon     6532    6534    .       +
unnamed-1       AUGUSTUS        initial 6532    6762    0.94    +       0
unnamed-1       AUGUSTUS        terminal        8714    10606   0.93    +
unnamed-1       AUGUSTUS        intron  6763    8713    0.89    +       .
unnamed-1       AUGUSTUS        CDS     6532    6762    0.94    +       0
unnamed-1       AUGUSTUS        CDS     8714    10606   0.93    +       0
unnamed-1       AUGUSTUS        exon    8714    10861   .       +       .
unnamed-1       AUGUSTUS        stop_codon      10604   10606   .       +
unnamed-1       AUGUSTUS        tts     10861   10861   .       +       .

# start gene g2
unnamed-1       AUGUSTUS        gene    12415   13535   0.06    -       .
unnamed-1       AUGUSTUS        transcript      12415   13535   0.06    -
unnamed-1       AUGUSTUS        tts     12415   12415   .       -       .
unnamed-1       AUGUSTUS        exon    12415   12983   .       -       .
unnamed-1       AUGUSTUS        stop_codon      12512   12514   .       -
unnamed-1       AUGUSTUS        terminal        12512   12983   1       -
unnamed-1       AUGUSTUS        initial 13169   13434   0.74    -       0
unnamed-1       AUGUSTUS        intron  12984   13168   1       -       .
unnamed-1       AUGUSTUS        CDS     12512   12983   1       -       1
unnamed-1       AUGUSTUS        CDS     13169   13434   0.74    -       0
unnamed-1       AUGUSTUS        exon    13169   13535   .       -       .
unnamed-1       AUGUSTUS        start_codon     13432   13434   .       -
unnamed-1       AUGUSTUS        tss     13535   13535   .       -       .
```

Figure 6: AUGUSTUS results

6532–6762 and 8714–10606. Gene 2, on the reverse strand, spans positions 12415–13535 and also contains two exons with an intron between them. The stop codon is in exon 1 (12512–12514), and the start codon is in exon 2 (13432–13434). The CDS includes two regions: 12512–12983 and 13169–13434. Both genes encode functional proteins. This detailed output highlights exon-intron boundaries, coding regions, and predicted protein sequences, which are valuable for downstream analyses like functional annotation and experimental validation.

DNA Subway AUGUSTUS

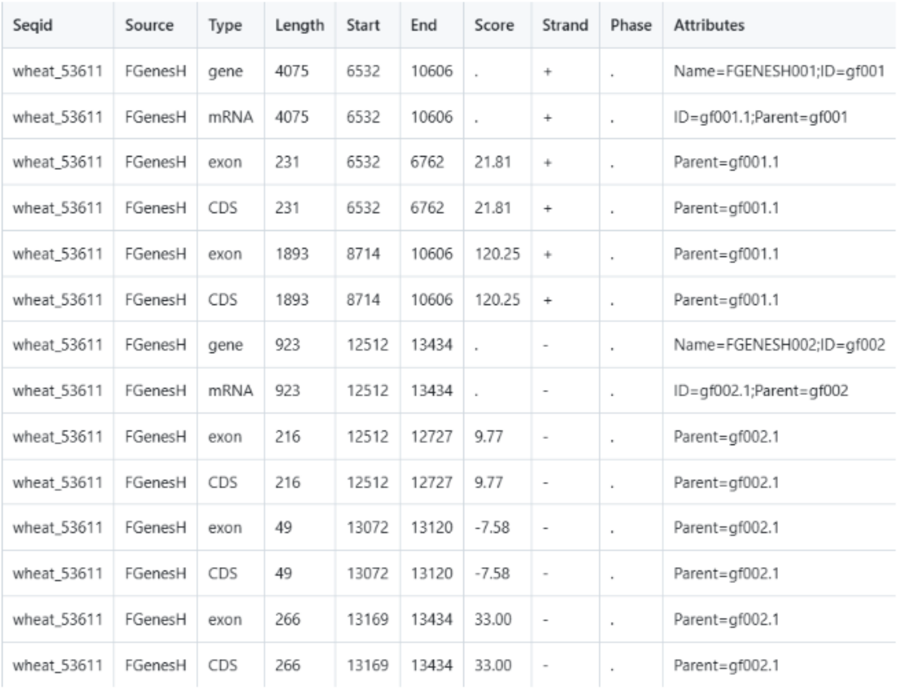


Seqid	Source	Type	Length	Start	End	Score	Strand	Phase	Attributes
wheat_53611	AUGUSTUS	gene	4075	6532	10606	0.87	+	.	Name=AUGUSTUS001;ID=g1
wheat_53611	AUGUSTUS	mRNA	4075	6532	10606	0.87	+	.	ID=g1.t1;Parent=g1
wheat_53611	AUGUSTUS	CDS	231	6532	6762	1	+	0	Parent=g1.t1
wheat_53611	AUGUSTUS	exon	231	6532	6762	1	+	0	Parent=g1.t1
wheat_53611	AUGUSTUS	CDS	25	8714	8738	0.87	+	0	Parent=g1.t1
wheat_53611	AUGUSTUS	exon	25	8714	8738	0.87	+	0	Parent=g1.t1
wheat_53611	AUGUSTUS	CDS	97	8824	8920	0.87	+	2	Parent=g1.t1
wheat_53611	AUGUSTUS	exon	97	8824	8920	0.87	+	2	Parent=g1.t1
wheat_53611	AUGUSTUS	CDS	1615	8992	10606	0.87	+	1	Parent=g1.t1
wheat_53611	AUGUSTUS	exon	1615	8992	10606	0.87	+	1	Parent=g1.t1
wheat_53611	AUGUSTUS	gene	929	12512	13440	0.54	-	.	Name=AUGUSTUS002;ID=g2
wheat_53611	AUGUSTUS	mRNA	929	12512	13440	0.54	-	.	ID=g2.t1;Parent=g2
wheat_53611	AUGUSTUS	CDS	472	12512	12983	0.88	-	1	Parent=g2.t1
wheat_53611	AUGUSTUS	exon	472	12512	12983	0.88	-	1	Parent=g2.t1
wheat_53611	AUGUSTUS	CDS	272	13169	13440	0.54	-	0	Parent=g2.t1
wheat_53611	AUGUSTUS	exon	272	13169	13440	0.54	-	0	Parent=g2.t1

Figure 7: DNA subway AUGUSTUS results

The AUGUSTUS tool identified two genes, g1 and g2, in the wheat sequence wheat_53611. Gene 1, located on the forward strand, spans positions 6532–10606 with a length of 4075 bp and a high prediction score of 0.87. It consists of four exons, with CDS regions ranging from 6532–6762, 8714–8738, 8824–8920, and 8992–10606. Gene 2, located on the reverse strand, spans positions 12512–13440 with a length of 929 bp and a prediction score of 0.54. It contains two exons, with CDS regions spanning 12512–12983 and 13169–13440. These predictions highlight the structural details of both genes, including exon-intron boundaries and coding sequences, which are critical for downstream analyses such as functional annotation and protein prediction.

DNA Subway FGENESH

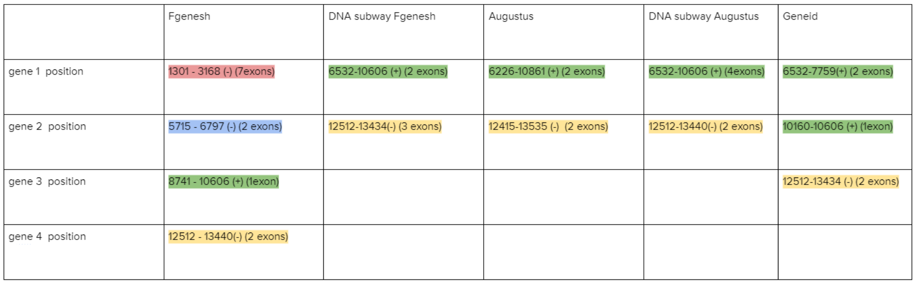


Seqid	Source	Type	Length	Start	End	Score	Strand	Phase	Attributes
wheat_53611	FGenesH	gene	4075	6532	10606	.	+	.	Name=FGENESH001;ID=gf001
wheat_53611	FGenesH	mRNA	4075	6532	10606	.	+	.	ID=gf001.1;Parent=gf001
wheat_53611	FGenesH	exon	231	6532	6762	21.81	+	.	Parent=gf001.1
wheat_53611	FGenesH	CDS	231	6532	6762	21.81	+	.	Parent=gf001.1
wheat_53611	FGenesH	exon	1893	8714	10606	120.25	+	.	Parent=gf001.1
wheat_53611	FGenesH	CDS	1893	8714	10606	120.25	+	.	Parent=gf001.1
wheat_53611	FGenesH	gene	923	12512	13434	.	-	.	Name=FGENESH002;ID=gf002
wheat_53611	FGenesH	mRNA	923	12512	13434	.	-	.	ID=gf002.1;Parent=gf002
wheat_53611	FGenesH	exon	216	12512	12727	9.77	-	.	Parent=gf002.1
wheat_53611	FGenesH	CDS	216	12512	12727	9.77	-	.	Parent=gf002.1
wheat_53611	FGenesH	exon	49	13072	13120	-7.58	-	.	Parent=gf002.1
wheat_53611	FGenesH	CDS	49	13072	13120	-7.58	-	.	Parent=gf002.1
wheat_53611	FGenesH	exon	266	13169	13434	33.00	-	.	Parent=gf002.1
wheat_53611	FGenesH	CDS	266	13169	13434	33.00	-	.	Parent=gf002.1

Figure 8: DNA Subway FGENESH

The FGENESH tool identified two genes, gf001 and gf002, in the wheat sequence wheat_53611. Gene gf001, located on the forward strand, spans positions 6532–10606 with a length of 4075 bp. It consists of two exons, the first spanning 6532–6762 (231 bp, score 21.81) and the second spanning 8714–10606 (1893 bp, score 120.25). Both exons contribute to the coding sequence (CDS). Gene gf002, located on the reverse strand, spans positions 12512–13434 with a length of 923 bp. It contains three exons: the first spans 12512–12727 (216 bp, score 9.77), the second spans 13072–13120 (49 bp, score -7.58), and the third spans 13169–13434 (266 bp, score 33.00). These detailed annotations provide insights into gene structures, exon positions, and strand orientation, making them valuable for downstream analysis and functional studies.

Gene Position	Fgenesh (exons)	DNA Subway Fgenesh (exons)	Augustus (exons)	DNA Subway Augustus (exons)	Geneid (exons)
Gene 1	1301–3168 (-) (7 exons)	6532–10606 (+) (2 exons)	6226–10861 (+) (2 exons)	6532–10606 (+) (4 exons)	6532–7759 (+) (2 exons)
Gene 2	5715–6797 (-) (2 exons)	12512–13434 (-) (3 exons)	12415–13535 (-) (2 exons)	12512–13440 (-) (2 exons)	10160–10606 (+) (1 exon)
Gene 3	8741–10606 (+) (1 exon)	-	-	-	-
Gene 4	12512–13440 (-) (2 exons)	-	-	-	12512–13434 (-) (2 exons)



	Fgenesh	DNA subway Fgenesh	Augustus	DNA subway Augustus	Geneid
gene 1 position	1301 - 3168 (-) (7exons)	6532-10606 (+) (2 exons)	6226-10861 (+) (2 exons)	6532-10606 (+) (4exons)	6532-7759(+) (2 exons)
gene 2 position	5715 - 6797 (-) (2 exons)	12512-13434(-) (3 exons)	12415-13535 (-) (2 exons)	12512-13440(-) (2 exons)	10160-10606 (+) (1exon)
gene 3 position	8741 - 10606 (+) (1exon)				12512-13434 (-) (2 exons)
gene 4 position	12512 - 13440(-) (2 exons)				

Figure 9: Table with colored labels for follow up

Common regions:

- **Gene 1** (green region) : The gene spanning from 6532 to 10606 on the forward strand (+), marked in green in the results of the DNA subway Fgenesh, is a consistent feature across multiple gene prediction tools but with some variations. Augustus and DNA subway augustus predict this gene at a slightly extended position from 6226 to 10861 with 2 exons for augustus and from 6532 to 10606 with 4 exons for DNA subway augustus. Geneid also identifies this region but divides it into two separate predictions: one from 6532 to 7759 (2 exons) and another one from 10160 to 10606 (1exon). This split in Geneid's prediction suggests a possible alternative structure or fragmentation. FGENESH, while differing in interpretation, may be representing the same gene with variation, as it predicts a single-exon gene extending from 8741 to 10606 (the green region) on the forward strand, aligning with the green region predicted in the other tools.

For the other region predicted by FGENESH that extends from 5715 to 6797 (the blue region) on the reverse strand (-) with 2 exons, it is not supported by the other tools. This region may partially overlap with predictions made by tools that focus on nearby regions but it does not appear explicitly as a standalone gene in the results of tools like AUGUSTUS or GENEID or DNA subway Fgenesh.

- **Gene 2** (yellow region) : the gene extending from 12512 to 13434 on the reverse strand (-) with 3 exons, marked in yellow in the results of the DNA subway Fgenesh, is a consistent feature across multiple gene prediction tools but with slight differences in exon count and exact positions. FGENESH predicts this region as a gene with 2 exons, extending from 12512 to 13440, aligning with the prediction of DNA subway Fgenesh. Augustus predicts this region as a gene spanning from 12415 to 13535 with 2 exons, slightly extending the boundaries compared to FGENESH. GENEID matches closely with FGENESH, predicting this gene at 12515 to 13434 with 2 exons. This consistency in identifying this region across tools indicates that it is a reliable gene prediction, with the variations in exon count and precise start-end positions reflecting the differences in each tool's algorithm.

Non common regions:

The region highlighted in *pink* (1301-3168 on the reverse strand) is a unique prediction made exclusively by FGENESH. According to FGENESh, this region has 7 exons starting from 1301 and ending at 3168. This prediction is not supported by any of the other tools used in the analysis, such as DNA Subway FGENESH, AUGUSTUS or Geneid which do not identify a gene in this specific region. The lack of agreement from other tools suggests that this region might be an artifact of the FGENESH algorithm, a false positive, or a region with characteristics that make it detectable only by FGENESH. Alternatively, it could represent a low-confidence or poorly conserved gene that is difficult for other tools to detect

$\Longleftrightarrow$ These variations between the tools highlight the need for further investigation, validation as through transcriptomic data, to confirm the existence, structure, and functionality of these predicted genes and to ensure biologically meaningful results.

> p.s. In order to visualize the features predicted by the abovementioned tools on artemis we first need to convert them to `.gff` format.

Gene Validation

BLAST

We will perform a BLAST search of the predicted genes against related species proteomes locally, using blast+ package on unix terminal.

Blasting done through `blastp` program on the predicted genes from

> region8 (each tool generated .faa fasta file that is amino acid sequences, each sequence will be a query).

The proteomes were retrieved from UniProt as it is recommended to be the way to download proteome for a whole species by an EMBL-EBI training course on UniProt[4], we will provide an api to retrieve the sequences as well to make our work replicable.

Since we're blasting against local databases built from proteomes retrieved from UniProt, the resulting hits have UniProt IDs, instead of product or gene names, thus an extra processing step was taken to annotate the results through a python script using `biopython`'s `ExPASy` and `SwissProt` modules.

> On another note, the advantages of blasting locally here on particular species is it's more specific and centered towards the species of interest, and provides a larger set of similar proteins in comparison to swissport for instance, which has a very limited number of reviewed proteins in each of the species we are interested in, which will be noted in the results.

Tritium aestivum proteome

To validate the predicted genes, we will start of by blasting against the proteome of *Triticum aestivum* available on UniProt. We retrieved the list of proteins from the supplementary material of an International Wheat Genome Sequencing Consortium (IWGSC) published in *Science*[5] aiming to provide an annotated reference sequence of the *Triticum aestivum* genome. The article is available here. We will access all the proteins sequences (including isoforms) using an api call to the UniProt database.

```
$ curl https://rest.uniprot.org/uniprotkb/stream?compressed=true&format=fasta&query=%28%28l
    > data/sequences/proteome/Triticum_aestivum_proteins.fasta.gz
$ gunzip data/sequences/proteome/Triticum_aestivum_proteins.fasta.gz
$ cat data/sequences/proteome/Triticum_aestivum_proteins.fasta | grep '>' | wc -l
130283
```

There is a total of 130,283 proteins in the file.

We create the database locally, in `data/database`:

```
# --creating the local database
# 1. Tritium_aestivum_proteome
makeblastdb -in data/web_retrieved_sequences/proteins.fasta \
            -dbtype prot \
            -out data/database/Triticum_aestivum_proteome/Triticum_aestivum_proteome
```

[4] EMBL-EBI training course on UniProt: https://www.ebi.ac.uk/training/online/courses/uniprot-exploring-protein-sequence-and-functional-info/

[5] The International Wheat Genome Sequencing Consortium (IWGSC) et al. ,Shifting the limits in wheat research and breeding using a fully annotated reference genome.Science361,eaar7191(2018).DOI:10.1126/science.aar7191

We will now perform a BLAST search against this database to see if our predicted genes are similar to any of the known annotated proteins of *Triticum aestivum*. It'll be a blastp search, as we are looking for protein sequences that are similar to our predicted sequence (which is already translated by our tools output and saved in out repository in .faa files)

```
# a. on AUGUSTUS_predicted.faa output
blastp -query output/AUGUSTUS/AUGUSTUS_predicted.faa \
       -db data/database/Triticum_aestivum_proteome/Triticum_aestivum_proteome \
       -out output/blast/tabulated/AUGUSTUS_Triticum_aestivum_proteome_results.txt \
       -outfmt 6
       #tabulated output
```

Then, using biopython uniprot api again, we will annotate the results to have a better understanding of the hits, as the proteome contain solely IDs and sequences. And for that we created a python script to clean the blast results and add the uniprot annotation to it, running it this way:

```
python src/clean_blast_results.py output/blast/tabulated/AUGUSTUS_Triticum_aestivum_proteome
```

We provided the commands to make blast databases and perform all the search we've done in out analysis in the `blast.sh` script.

```
./src/blast.sh # to run all the blast commands
```

Related species

Starting from the following information:
- *Triticum monococcum* and *Triticum durum* have the A and B chromosomes
- *Aegilops tauschii* has the D chromosome

We will also look for their proteomes and perform the same blasting procedure as above.

N.B: we couldn't find *Tri. monococcum* proteome on UniProt, so we will only blast against *Tri. durum* for the common A and B chromosomes.

Triticum durum The proteome can be find on this UniProt page, 188,826 proteins, worth noting that only 2 of them are expertly reviewed - Swiss-Prot - the rest are unreviewed - TrEMBL.

For *Triticum durum*, retrieving the proteome through this api call:

```
$ curl https://rest.uniprot.org/uniprotkb/stream?compressed=true&format=fasta&query=%28%28ta
    > data/sequences/proteome/Triticum_durum_proteins.fasta.gz
$ gunzip data/sequences/proteome/Triticum_durum_proteins.fasta.gz
```

Aegilops tauschii The proteome can be find on *this UniProt page*, 214,193 proteins, only one of them is expertly reviewed.

Retrieving the proteome through this api call:

```
$ curl https://rest.uniprot.org/uniprotkb/stream?compressed=true&format=fasta&query=%28%28ta
    > data/sequences/proteome/Aegilops_tauschii_proteins.fasta.gz
$ gunzip data/sequences/proteome/Aegilops_tauschii_proteins.fasta.gz
```

Results

In this results section we are, as explained, expecting to have for each predicted gene 3 blasting results. Since we're taking into consideration 4 genes from FGENESH and 2 from AUGUSTUS and blasting against 3 species' proteomes separately (T. aestivum, A. tauschii and T. durum), we would have 6 genes to analyse with 3 resulting blast output each. *BLAST results can be found by clicking on*: this link
In the first reported blast we will analyse every single detail extensively to give an intuition of our analysis

Augustus gene 1: (*protein length: 707aa*) In the first BLAST results for AUGUSTUS against our own species's proteome *Triticum aestivum*:
We find the top 5 hits quite significant with % id higher than 98.7% and then immediately drops to 52% after these 5 matches from the proteome database, providing an e-value estimated by blast+ to be 0 which is quite significant, thus we will be considering them as top results. Looking into them, we find the 1st hit to be 100% id, found on chromosome 4D, matching all of the protein's length (1-707 residues) against the full length of the subject from the database (also 1-707) with NO mismatches NO gaps, and this exact match gives a product: *Anaphase-promoting complex subunit 11 {ECO:0008006|Google:ProtNLM}, with gene name CFC21_063427 (as retrieved from uniprot's api)*

a snippet of the blast results and retrieved information for this hit:

- **Transcript ID:** AUGUSTUS_g1.t1
- **Protein ID:** tr|A0A3B6JR29|A0A3B6JR29_WHEAT
- **Alignment Score:** 100
- **Query Length:** 707
- **Mismatch Count:** 0
- **Gap Count:** 0
- **Query Start:** 1
- **Query End:** 707
- **Subject Start:** 1
- **Subject End:** 707
- **E-value:** 0
- **Bit Score:** 1462
- **Protein Name:** RecName: Full=Anaphase-promoting complex subunit 11 {ECO:0008006|Google:ProtNLM}
- **Chromosome:** Chromosome 4D
- **Organism:** Triticum aestivum (Wheat)
- **Organism Protein ID:** A0A3B6JR29_WHEAT
- **ORF Names:** CFC21_063427_063427 {ECO:0000313|EMBL:KAF7055962.1}

- **Keywords:** Metal-binding, Reference proteome, Zinc, Zinc-finger

All the hits follows have the same product name, same gene name and aroudn same length (707 or 708 due to inserted gap). We can notice that not all of them are on the several hits can be due to:

- duplication
- isotopes (not an expertly reviewed database liek swissprot)
- hits on the same protein sequence but different alignments patterns

Having the match 100% id to the first protein is a validation besides all the above mentioned signs from results (consistency of the matched proteins among the best hit), meaning that this gene moght be infat the CFC21_063427 gene, and the protein is the Anaphase-promoting complex subunit 11.
More interestingly it resides on the chromosome 4D, which is the chromosome we have mapped our region to, which provide a stringent evidence of our corelated work.

The 2nd blast is done on *Aegilops tauschii* proteome:

we find the 1st hit to be 100% id, matching all of the protein's length (1-707 residues) against the full length of the subject from the database (also 1-707) with only 5 mismatches, and this exact match gives a product: *Anaphase-promoting complex subunit 11 {ECO:0008006|Google:ProtNLM} (as retrieved from uniprot's api)*

The 3 hits that follows are truncated to be around 600 residues of length (630, 629, 621 respectively) and 4 mismatches beginning at around 78-87th position of the query until the end, each corresponding to these proteins: *Anaphase-promoting complex subunit 11 (twice) and VWFA domain-containing protein {ECO:0008006|Google:ProtNLM}*.

And the last hit is strictly a small segment from the first 388 residues of the query corresponding to *RING-type domain-containing protein {ECO:0008006|Google:ProtNLM}*, which can be a suggesting that the first part of the query contains this particular domain.

Even though the annotations are quite different, 2 interesting things are worth noting: All of them are metal-binding domains containing proteins relating to zinc and zinc-fingers, as can be seen from the list of keywords extracted from uniprot from the hit id, thus showing even though there is no consensus towards the annotation there is majority agreement on the functionality and domains of this sequence product:

> N.B. proteins all show to be on the same segment (chromosome), now it's worth noting that these are not expertly annotated as found in databases like swiss port, but overall these proteins are very highly similar between each other too.

Next one with *Triticum durum* proteome, which only has A and B chromosome (and in our case we have a D chromosome) but also showed 3 significant results belonging to the same protein (CFC21_063427 gene), the 1st 2 of the same length belonging to different chromosomes providing evidence to presence of duplicates maybe.

On a side note:

- the fact that the same protein is found in 4D of our species and 4D of *Aegilops tauschii* enforces an evolutionary interspecies link (also in our region which turned out to be on 4D, this is a bonus finding)

- It is also perceived in the other species, *Triticum durum*, that only has the A and B which is a sign of conservation of this protein among the genus of *Triticum*, implying its importance maybe in this organism.

All these blast results show that indeed this gene1 is highly likely to be associated with the *Anaphase-promoting complex subunit 11* protein, belonging to the 4D chromosome of our species, and there exist general agreement of the use of is full length (707aa) and is associated with a zinc-finger motif along with metal-binding activity (zinc most probably), with 100% match to the same protein product from our own species. We shall see other results to see the validity of our assumptions

Augustus gene 2: (*protein length: 245aa*) *Against Triticum aestivum proteome*

Results firstly show 100% match of the first 245 residues of *Uncharacterized protein {ECO:0000313|EMBL:KAF7055961.1, ECO:0000313|EnsemblPlants:TraesCS4D02G339100.1}*, activity in DNA binding and transcription regulation, localized subcelluraly in the nucleus which might indicate its possibility to be a transcription factor. The aligned subject starts at 3rd residue, protein of length 247 indicating a possible mistake in the prediction of the gene length by AUGUSTUS.

The other 2 hits also have high % match (>95), same query length and position mapped but differ only in the start position of the subject protein and the chromosome location (do not map in 4D) suggesting possible duplicates.

Aegilops tauschii

Only one significant hit, same position as found previously 1-245, same exact annotation of the *Uncharacterized protein* with transcription involving function found on chromosome 4D of this species (showing evolutionary conservation).

Triticum durum

Same results almost, same positions are aligned (1-245), same functional annotation, compartmentalization (in nucleus) for 2 hits on respectively 5b and 4a chromosomes (near 4d position wise on whole genome). Also this uncharacterized protein provides evidence on the structural annotation of these 245 residues of our genomic region done by AUGUSTUS

Thus the 2nd prediction of AUGUSTUS is also highly likely to be validated, with provided proof on conservative 245 residues involved in dna binding activity

Fgenesh gene 1 & 2 Show absolutely no significant results in each of the species, one of them no hits at all, those that have have a higher proportion mismatches than matches, high e value and % id <40. Thus no validation of these gene structures predicted by Fgenesh

Fgenesh gene 3 (length: 621aa) This one provides interestingly similar results that also enforce our previously established hypothesis in the first Augustus gene:

Triticum aestivum

5 hits above 92%, all of them against the same protein (the *Anaphase-promoting complex subunit 11*)
Localized on chromosome 4D, matching the same length of our query (make sure) 1-621 in all 5 hits to different positions of the subject starting either at 67th or 76th residue with different alignment gaps and mismatch patterns

> a more subtle hypothesis then having the same exact protein of different versions in the same place is having it align to the same protein entry, but we can not assume that especially that each of these 5 proteins has a different UniProt ID

Aegilops tauschii

These show similarity to AGUSTUS gene 1 results against this species' proteome. 4 hits of exactly 99.345% identity, the length is similar to previously reported (1-621 of query except the last one is 1-620) showing matches to respectively: *VWFA domain-containing protein* & *Anaphase-promoting complex subunit 11* (x2)
Also all metal bonding activity and present on 4D chromsome of this genome (matches with Augustus prediction)

Triticum durum

These results show top 3 >92% id with Anaphase-promoting complex subunit 11, also 1-621 against 87-707 of 2 subjects and 87-670 of the 3rd one. Thus promoting the same resulting conclusion

For this gene Fgenesh3, we would like to refer back to the analysis for Augustus gene 1 as something interesting is happening:

As previously stated from gene prediction results, these 2 predicted genes might be attempts to map the same gene, which is validated here through their mapping of Anaphase.. with high %id emphasising this segment's functionality

we can notice that August's prediction match the 1-707 of subject while Fgenesh start at either 67 or 76th residue of the subject, both matching the length of the predicted genes length (707 vs 621 residues), we can comment that Fgensh has a truncated prediction at the 5' UTR which is supposed to be part of the product. (We can say that the 621 aa residues have been well predicted in the coding region, there might be a truncated chunk at the beginning of the protein as predicted by Fgenesh as the matches are late to start on subject proteins and when put in comparison with the length and alignment of Augustus's)

Fgenesh gene 4 (length: 247aa)

Finally, we have here in *Triticum aestivum* proteome, 3 hits with 100% id, 247 residues of the query matching the same length of the subject, all of them are *Uncharacterized protein* but with same identifiers involved with ECO and such (also identifiers relatd to DNA-binding) with the same functional annotation, teh 100% match is with subject on 4D teh other 2 are respectively 4B and 5A which are closest to 4D in position, with the same exact alignment pattern, no mismatches, no gaps, and the same exact length of the protein.

2 further hits in *Triticum durum* and 1 in *Aegilops tauschii* show the same results.

Worth noting this matches with subject protein's length and further validates Augustus results, but even better with teh 2 first missing residues that we have mentioned before present here, thus highly suggesting the presence of a coding region here.

Transcriptome

Testing if we can find any transcriptome data

The *European Nucleotide Archive (ENA)* comprises a large collection of sequencing data from raw sequences to assembly to functionally annotated ones. While looking for transcriptome studies for *Triticum aestivum* we find several projects

(Total= 22, in this table[6])

TSA stands for Transcriptome Shotgun Assembly

One of them is published by Xiao et al. (2013) in BMC Genomics [7]. They have performed short read RNA-seq using Illumina Hi-Seq tech, and deposited the project's raw reads on the SRA database, project `SRX212270`. We will use this as trial to explore how we can validate using Whole Transcriptomes before optimizing our choice. While running out of time and memory, we will try doing that using Galaxy[8][9].

Working on galaxy, first retrieve the SRA accession number from the project, tools > Get data > EBI SRA, copy the accession number and get the fastq in galaxy. After loading them (paired end so 2 fastq) > fastq groomer, to make sure the fastq format fits Galaxy's requirement and make it run. Meanwhile > FastQC to make sure the quality of the transcriptome is good or whether it's

[6]

Accession	Description
GAEF01000000	Triticum aestivum, TSA project GAEF01000000 data
GAJL01000000	Triticum aestivum, TSA project GAJL01000000 data
GBKH01000000	Triticum aestivum, TSA project GBKH01000000 data
GBKI01000000	Triticum aestivum, TSA project GBKI01000000 data
GBKJ01000000	Triticum aestivum, TSA project GBKJ01000000 data
GBKK01000000	Triticum aestivum, TSA project GBKK01000000 data
GBZP01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.
GDTJ01000000	Triticum aestivum, TSA project GDTJ01000000 data
GEUX01000000	Triticum aestivum, TSA project GEUX01000000 data
GEWU01000000	Triticum aestivum, TSA project GEWU01000000 data
GFFI01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.
GIJS01000000	Triticum aestivum, TSA project GIJS01000000 data
GILY01000000	Triticum aestivum, TSA project GILY01000000 data
GIXT01000000	TSA: Triticum aestivum cultivar TcLr19 isolate leaf, transcriptome shotgun assembly.
GJAR01000000	TSA: Triticum aestivum cultivar Avocet R, transcriptome shotgun assembly.
GJUY01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.
HAAB01000000	Triticum aestivum, TSA project HAAB01000000 data
HCEC01000000	TSA: Triticum aestivum
HCED01000000	TSA: Triticum aestivum
IAAK01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.
IAAL01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.
IAAM01000000	TSA: Triticum aestivum, transcriptome shotgun assembly.

[7] Xiao, J., Jin, X., Jia, X., Wang, H., Cao, A., Zhao, W., . . . & Wang, X. (2013). Transcriptome-based discovery of pathways and genes related to resistance against Fusarium head blight in wheat landrace Wangshuibai. BMC genomics, 14, 1-19.

[8] The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update Nucleic Acids Research, gkae410 doi:10.1093/nar/gkae410

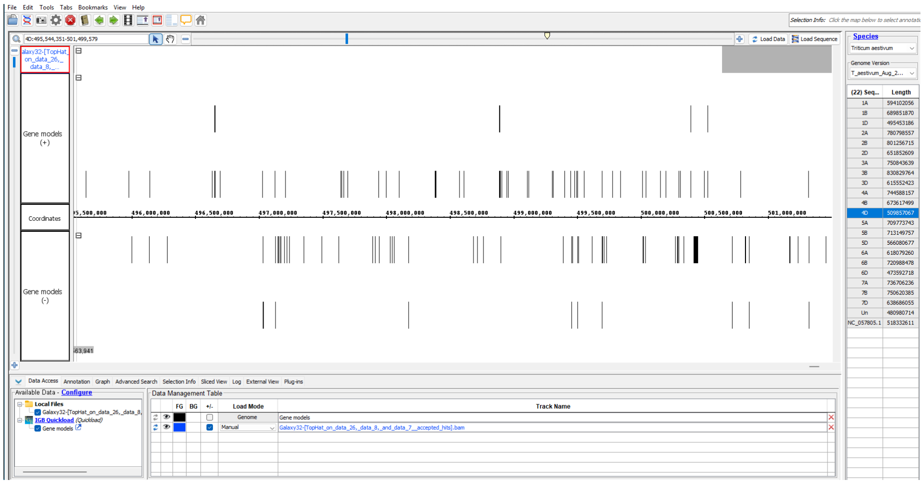
[9] The Galaxy server used for some calculations is partly funded by the German Federal Ministry of Education and Research BMBF grant 031 A538A de.NBI-RBC and the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) within the framework of LIBIS/de.NBI Freiburg.

better to take another set of reads.

We will try now mapping: using Tophat2, we will map the reads to the reference genome of *Triticum aestivum* (available on ENSEMBL) to see how many reads are mapped and how many are not. We have taken the reference genome using

trying to perform RNA-seq aln and viz using IGB, no reads show

We also tried performing mapping through RNA Star, which is a splice aware and fast performing aligner, but the results were not satisfactory, also no reads were shown in the region of interest.



Species: Triticum aestivum

Genome Version: T_aestivum_Aug_2...

(22) Seq...	Length
1A	594102056
1B	689851870
1D	495453186
2A	780798557
2B	801256715
2D	651852609
3A	750843639
3B	830829764
3D	615552423
4A	744588157
4B	673617499
4D	509857067
5A	709773743
5B	713149757
5D	566080677
6A	618079260
6B	720988478
6D	473592718
7A	736706236
7B	750620385
7D	638686055
Un	480980714
NC_057805.1	518332611

Data Management Table

FG	BG	+/-	Load Mode	Track Name
■		☐	Genome	Gene models
■		☑	Manual	Galaxy32-[TopHat_on_data_26,_data_8,_and_data_7__accepted_hits].bam

Figure 10: Tophat2 resulting bam file viz: no reads are shown in our region

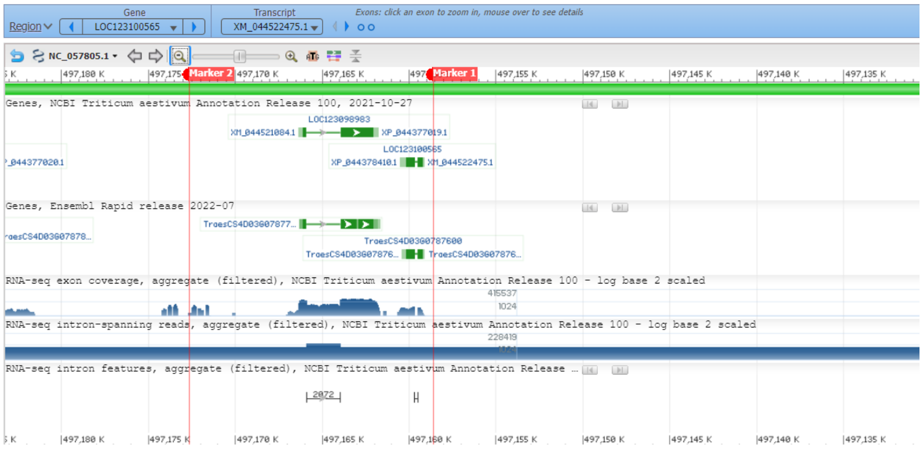
After all, we have used over 80GB of memory for over 60 hours of computing on Galaxy server, we were in fact no longer able to proceed with RNAseq analysis.

On the other hand we found this on NCBI, when we mapped our region to its coordinates on teh reference of chromosome 4D, and we found the following:

One transcript was shown with some exon RNAseq exon density in the region, providing some hope that this region can actually contain a gene, even though the transcript does not map exact locations that we have. But overall, this is a good sign that the region is transcribed and can contain a gene (or more).

Another trial:

cDNA (complementary DNA) is a single-stranded DNA synthesized from a messenger RNA (mRNA) template in a reaction catalyzed by the enzyme reverse transcriptase. It is thus synthesized from the mRNA template, it can be used to study the gene expression in a cell, as it is a copy of the mRNA, and can be used to study the gene expression in a cell. It's a representation of a gene's transcript. On Ensembl Plants, we can find the cDNA of *Triticum aestivum here on this*



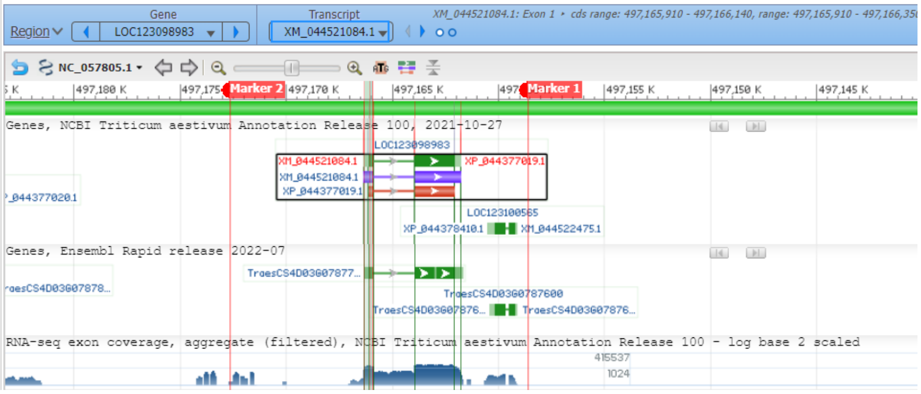
Gene: LOC123100565

Transcript: XM_044522475.1

Exons: click an exon to zoom in, mouse over to see details

Region
NC_057805.1
Marker 2
Marker 1
497,180 K
497,175 K
497,170 K
497,165 K
497,160 K
497,155 K
497,150 K
497,145 K
497,140 K
497,135 K
Genes, NCBI Triticum aestivum Annotation Release 100, 2021-10-27
LOC123098983
XM_044521084.1
XP_044377019.1
LOC123100565
XP_044378410.1
XM_044522475.1
P_044377020.1
Genes, Ensembl Rapid release 2022-07
TraesCS4D03G07877...
raesCS4D03G07878...
TraesCS4D03G0787600
TraesCS4D03G07876...
TraesCS4D03G07876...
RNA-seq exon coverage, aggregate (filtered), NCBI Triticum aestivum Annotation Release 100 - log base 2 scaled
415537
1024
RNA-seq intron-spanning reads, aggregate (filtered), NCBI Triticum aestivum Annotation Release 100 - log base 2 scaled
228419
1024
RNA-seq intron features, aggregate (filtered), NCBI Triticum aestivum Annotation Release ...
2072

Figure 11: region8 marked on chr 4D with RNAseq density, NCBI GDV



Gene: LOC123098983

Transcript: XM_044521084.1

XM_044521084.1: Exon 1 ‣ cds range: 497,165,910 - 497,166,140, range: 497,165,910 - 497,166,35

Figure 12: region8 marked on chr 4D with XM_044522475.1 transcript showing, NCBI GDV

ftp site (click link). There is one fasta file containing all of the genome's cDNA sequences, with a particular header format. To make the process more easily computable, we wrote a bash script to filter the cDNA sequences of the chromosome 4D (can be found in `./src/filter_cDNA.sh`) , and save them in a separate file. After that, we retrieved only 2 cDNA sequences:

```
>TraesCS4D02G339400.1 cdna chromosome:IWGSC:4D:497165754:497169019:-1
AGCCCCACCCATTTCCTTCCCTTCGGTCGAGGAAGGCAGCAGCAATAAATCTAGGTCCGG
>TraesCS4D02G339300.1 cdna chromosome:IWGSC:4D:497150642:497160941:-1
CTTCAAGAGATGGAGATCCCTGACCAGCAGCCTGCGGTCGCAGTCGCAGAGATGGAAGCC
```

Also showing an output out of the studied region, even though none of them directly validates any of our predictions.

> We tried looking for ESTs too but couln't find, however we did not try as extensively as RNA this is why it's not mentioned

Transposable Elements (TEs)

To detect transposable elements (TEs) in the genomic region of *Triticum aestivum* (wheat), we used the following tools:

- RepeatMasker software with the TREP database considering first the Triticum genus only and then the complete database.
- Censor from Genetic Information Research Institute (GIRI) website, considering first the Triticum genus and then the Viridiplantae.
- RepeatMasker included in DNA Subway.

RepeatMasker

RepeatMasker Website

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. It can be used to identify transposable elements (TEs) in genomic sequences.

Prerequisites

1. **Perl**:

   Verify that Perl version 5.8.0 or higher is installed: `bash  perl -v` If not installed, run: `bash  sudo apt update  sudo apt install perl`

2. **Python 3 and h5py Library**:

   Verify that Python 3 is installed: `bash  python3 --version` Then, install the `h5py` library: `bash  sudo apt install h5py`

3. **Sequence Search Engine: RMBlast**:

We will use **RMBlast**, a RepeatMasker-specific version of NCBI BLAST, which is optimized for repeat detection and recommended for use with RepeatMasker, particularly for complex genomes like *Triticum aestivum* (RMBlast Website).

Download RMBlast: `bash  wget https://www.repeatmasker.org/rmblast/rmblast-2.14.1+-x64-li` Extract and move it to the system's PATH: `bash  tar zxvf rmblast-2.14.1+-x64-linux.tar.gz  sudo mv rmblast-2.14.1 /usr/local/bin/rmblast` Remove the downloaded tar file: `bash  rm rmblast-2.14.1+-x64-linux.tar.gz`

4. **Tandem Repeat Finder (TRF)**:

   Download TRF: `bash  wget https://github.com/Benson-Genomics-Lab/TRF/releases/download/v4` Make the file executable and move it to the system's PATH: `bash  chmod +x trf409.linux64  sudo mv trf409.linux64 /usr/local/bin/trf`

Installation

1. **Download RepeatMasker**:

   The latest version is **RepeatMasker-4.1.7-p1.tar.gz**. To download the file in the /usr/local directory: `bash  cd /usr/local/  sudo wget https://www.repeatmasker.org/RepeatMasker/RepeatMasker-4.1.7-p1.tar.gz`

2. **Unpack the Distribution**:

   Unpacking it to `/usr/local/` directory: `bash  sudo gunzip RepeatMasker-4.1.7-p1.tar.gz  sudo tar xvf RepeatMasker-4.1.7-p1.tar sudo rm RepeatMasker-4.1.7-p1.tar`

3. **Run the Configure Script**:

   ```
   cd /usr/local/RepeatMasker
   sudo perl ./configure
   ```

4. **Install RepeatMasker Libraries**:

   We installed Dfam Viridiplantae partition database but it required a lot of space and time to use. Also, for the RepBase database it requires subscription. So we will use the **TREP Database**.

   TREP Database:

   TRansposable Elements Platform (TREP) is a curated collection of transposable elements (TEs) originally focused on **Triticeae species** (wheat, barley, maize), but has expanded to include TEs from various other species. This database is essential for identifying, classifying, and masking TEs in genomic sequences (TREP Database).

   To download the necessary TREP database files:

1. **Triticum sequences**
   This database contains sequences specific to *Triticum* genus. `bash`
   `sudo wget https://trep-db.uzh.ch/blast/dir_download/sequences.zip -P /usr/local/RepeatMasker/Libraries/`
2. **Complete TREP nucleotide sequence database (4,162 sequences)**
   This database includes all TE entries for detailed analysis. `bash`
   `sudo wget https://trep-db.uzh.ch/downloads/trep-db_complete_Rel-19.fasta.gz -P /usr/local/RepeatMasker/Libraries/`

To extract the files:

```
cd /usr/local/RepeatMasker/Libraries/
sudo unzip sequences.zip
sudo gunzip trep-db_complete_Rel-19.fasta.gz
```

Usage

1. **RepeatMasker with Triticum sequences from TREP database:**

```
RepeatMasker -lib /usr/local/RepeatMasker/Libraries/sequences.fasta \
            -dir /home/joelle/M1/Structural_Genomics/triticum_sequences \
            /home/joelle/M1/Structural_Genomics/region8.fasta.txt
```

`-lib`: Specifies the library file to use for masking.
`-dir`: Specifies the output directory for the results.
`-region8.fasta.txt`: The input file containing the genomic sequence to analyze.

The output file region8.fasta.txt.out contains the following information:

SW score	perc div.	perc del.	perc ins.	query sequence	position in query (begin - end)	matching repeat	repeat class/family	position in repeat (begin - end)	ID
406	33.0	7.2	1.4	region8	129 - 533	RLC_Taes_Ida_EF540325-1	Unspecified	3537 - 3964 (126)	1
16	0.0	0.0	0.0	region8	2655 - 2671	(T)n	Simple_repeat	1 - 17 (0)	2
229	0.0	0.0	0.0	region8	3291 - 3316	DTT_Taes_SBB_42j2-2	Unspecified	9 - 34 (51)	3
811	3.1	0.0	0.0	region8	3736 - 3832	DTT_Taes_Hades_42j2-4	Unspecified	1 - 97 (0)	4
1032	30.8	3.9	4.5	region8	5201 - 5809	DTH_Taes_Rong_AY97092-1	Unspecified (C)	1045-1716 (1111)	5

SW score	perc div.	perc del.	perc ins.	query sequence	position in query (begin - end)	matching repeat	repeat class/family	position in repeat (begin - end)	ID
1060	33.8	1.7	1.5	region8	5333 - 5915	DTH_Tmon_unnamed_AY924380-1	Unspecified (C)	7 - 914 (660)	6
14	12.8	5.9	0.0	region8	6393 - 6426	(CTCC)n	Simple_repeat	1 - 36 (0)	7
18	13.3	0.0	0.0	region8	6488 - 6520	(GGGTC)n	Simple_repeat	1 - 33 (0)	8
15	0.0	4.5	0.0	region8	6618 - 6639	(CGC)n	Simple_repeat	1 - 23 (0)	9
827	3.1	0.0	0.0	region8	7432 - 7529	DTT_Taes_Icarus_BJ274200-1	Unspecified	5 - 102 (1)	10
13	10.8	0.0	9.1	region8	12248 - 12283	(TGGTCA)n	Simple_repeat	1 - 33 (0)	11
12	8.0	7.4	0.0	region8	12329 - 12355	(GCAT)n	Simple_repeat	1 - 29 (0)	12

- **SW score**: Smith-Waterman score.
- **perc div.**: Percentage of substitutions in the alignment.
- **perc del.**: Percentage of deletions in the alignment.
- **perc ins.**: Percentage of insertions in the alignment.
- **query sequence**: Name of the query sequence.
- **position in query (begin - end)**: Position of the alignment in the query sequence.
- **matching repeat**: Name of the matching repeat.
- **repeat class/family**: Class and family of the matching repeat.
- **position in repeat (begin - end)**: Position of the alignment in the repeat sequence.
- **ID**: Unique identifier for the alignment.

2. **RepeatMasker with the complete TREP database**:

```
RepeatMasker -lib /usr/local/RepeatMasker/Libraries/trep-db_complete_Rel-19.fasta \
            -dir /home/joelle/M1/Structural_Genomics/trep-db_complete_Rel-19 \
            /home/joelle/M1/Structural_Genomics/region8.fasta.txt
```

The output file region8.fasta.txt.out:

SW score	perc div.	perc del.	perc ins.	query sequence	position in query (begin - end)	matching repeat	repeat class/family	position in repeat (begin - end)	ID
912	26.2	13.8	1.3	region8	63 - 533	RLC_Hvul_Ikya_EF067844-1	Unspecified	5056 - 5584 (1177)	1 *
7408	13.2	1.1	6.7	region8	129 - 1537	RLC_Hvul_Ikeros_AY053246-1	Unspecified	5036 - 6390 (139)	2
676	24.4	4.5	2.5	region8	1299 - 1567	RLC_Hvul_Ikya_EF067844-1	Unspecified	1 - 274 (6487)	3 *
16	0.0	0.0	0.0	region8	2655 - 2671	(T)n	Simple_repeat	1 - 17 (0)	4
267	2.9	0.0	0.0	region8	3291 - 3324	DTT_Hvul_SBB_consensus-1	Unspecified	9 - 42 (47)	5
811	3.1	0.0	0.0	region8	3736 - 3832	DTT_Taes_Hades_42j21-4	Unspecified	1 - 97 (0)	6
393	22.1	7.8	5.3	region8	3738 - 3904	DTC_Atau_Jorge_A_consensus-1	Unspecified (C)	11255 - 11209 (11039)	7 *
1295	34.7	2.2	3.1	region8	5033 - 5882	DTH_Bdis_Bdis_C_consensus-1	Unspecified (C)	317 - 1655 (813)	8
14	12.8	5.9	0.0	region8	6393 - 6426	(CTCC)n	Simple_repeat	1 - 36 (0)	9
18	13.3	0.0	0.0	region8	6488 - 6520	(GGGTC)n	Simple_repeat	1 - 33 (0)	10
15	0.0	4.5	0.0	region8	6618 - 6639	(CGC)n	Simple_repeat	1 - 23 (0)	11
261	26.9	11.1	0.7	region8	7423 - 7557	DHH_Bdis_AA_Bd1-3	Unspecified (C)	563 - 12064 (11916)	12 *
827	3.1	0.0	0.0	region8	7432 - 7529	DTT_Taes_Icarus_BJ254200-1	Unspecified	1 - 102 (1)	13
243	11.8	11.7	1.5	region8	8092 - 8151	RIX_Hvul_Morpheus_AY504843-1	Unspecified	7443 - 7515 (451)	14
13	10.8	0.0	9.1	region8	12248 - 12283	(TGGTCA)n	Simple_repeat	1 - 33 (0)	15
12	8.0	7.4	0.0	region8	12329 - 12355	(GCAT)n	Simple_repeat	1 - 29 (0)	16

Censor

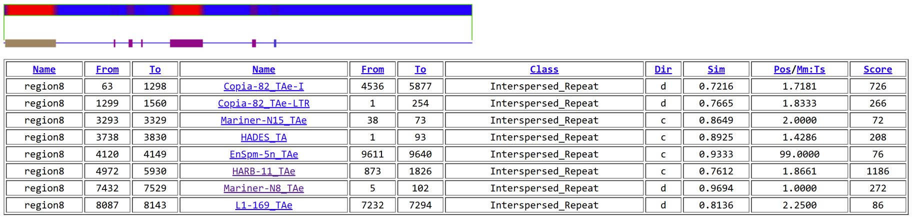
Censor website

Censor is a tool provided by the Genetic Information Research Institute (GIRI) that screens DNA sequences for interspersed repeats and low complexity DNA sequences. It uses Repbase, a database of repetitive DNA elements, to identify transposable elements (TEs) in genomic sequences.

1. **Censor with Sequence source set to Triticum genus**:

   The output can be found here: Censor Triticum genus output

   SVG Plot and table output:



Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
region8	63	1298	Copia-82_TAe-I	4536	5877	Interspersed_Repeat	d	0.7216	1.7181	726
region8	1299	1560	Copia-82_TAe-LTR	1	254	Interspersed_Repeat	d	0.7665	1.8333	266
region8	3293	3329	Mariner-N15_TAe	38	73	Interspersed_Repeat	c	0.8649	2.0000	72
region8	3738	3830	HADES_TA	1	93	Interspersed_Repeat	c	0.8925	1.4286	208
region8	4120	4149	EnSpm-5n_TAe	9611	9640	Interspersed_Repeat	c	0.9333	99.0000	76
region8	4972	5930	HARB-11_TAe	873	1826	Interspersed_Repeat	c	0.7612	1.8661	1186
region8	7432	7529	Mariner-N8_TAe	5	102	Interspersed_Repeat	d	0.9694	1.0000	272
region8	8087	8143	L1-169_TAe	7232	7294	Interspersed_Repeat	d	0.8136	2.2500	86

Figure 13: output1

   - **Name**: The name of the genomic region or sequence being analyzed.
   - **From**: The starting position of the sequence in the input genomic region.
   - **To**: The ending position of the sequence in the input genomic region.
   - **Name (Repeat)**: The name of the identified repeat element within the sequence.
   - **From (Repeat)**: The starting position of the repeat sequence in the genomic region.
   - **To (Repeat)**: The ending position of the repeat sequence in the genomic region.
   - **Class**: The type or class of the repeat element (e.g., Interspersed_Repeat, Simple_repeat).
   - **Dir**: The orientation of the repeat element relative to the genomic sequence (‘d’ for direct, ‘c’ for complementary).
   - **Sim**: The similarity score between the repeat sequence and the genomic region, indicating the match quality.
   - **Pos/Mm:Ts**: The positional or match/mismatch score, reflecting the alignment quality between the repeat and the sequence.
   - **Score**: A cumulative score indicating the strength or confidence level of the match between the repeat and the genomic region.

   The similarity scores are quite high, indicating a strong match between the identified repeat elements and the genomic sequence.

   The summary of the different classes of repeat elements identified in the genomic region:

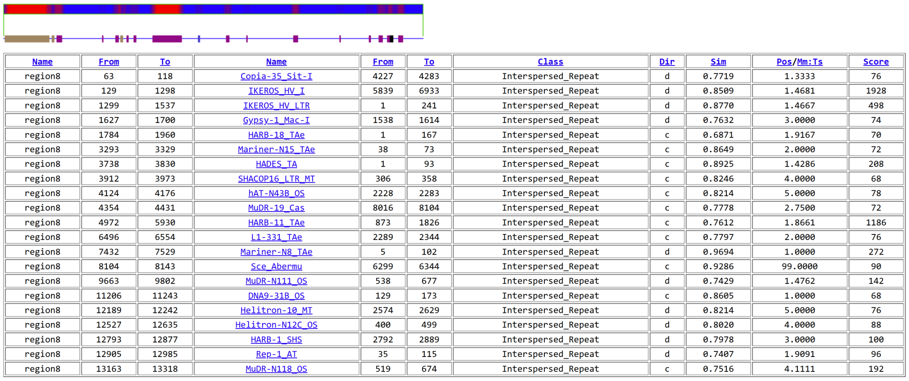
2. **Censor with Sequence source set to Viridiplantae**:

```
Repeat Class                                                    Fragments      Length
Transposable Element                                                    8        2772
   DNA transposon                                                       5        1217
      EnSpm/CACTA                                                       1          30
      Harbinger                                                         1         959
      Mariner/Tc1                                                       3         228
   LTR Retrotransposon                                                  2        1498
      Copia                                                             2        1498
   Non-LTR Retrotransposon                                              1          57
      L1                                                                1          57
Total                                                                   8        2772
```

Figure 14: output2

The output can be found here: Censor Viridiplantae output.

SVG Plot and table output:



Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
region8	63	118	Copia-35_Sit-I	4227	4283	Interspersed_Repeat	d	0.7719	1.3333	76
region8	129	1298	IKEROS_HV_I	5839	6933	Interspersed_Repeat	d	0.8509	1.4681	1928
region8	1299	1537	IKEROS_HV_LTR	1	241	Interspersed_Repeat	d	0.8770	1.4667	498
region8	1627	1700	Gypsy-1_Mac-I	1538	1614	Interspersed_Repeat	d	0.7632	3.0000	74
region8	1784	1960	HARB-18_TAe	1	167	Interspersed_Repeat	c	0.6871	1.9167	70
region8	3293	3329	Mariner-N15_TAe	38	73	Interspersed_Repeat	c	0.8649	2.0000	72
region8	3738	3830	HADES_TA	1	93	Interspersed_Repeat	c	0.8925	1.4286	208
region8	3912	3973	SHACOP16_LTR_MT	306	358	Interspersed_Repeat	c	0.8246	4.0000	68
region8	4124	4176	hAT-N43B_OS	2228	2283	Interspersed_Repeat	c	0.8214	5.0000	78
region8	4354	4431	MuDR-19_Cas	8016	8104	Interspersed_Repeat	c	0.7778	2.7500	72
region8	4972	5930	HARB-11_TAe	873	1826	Interspersed_Repeat	c	0.7612	1.8661	1186
region8	6496	6554	L1-331_TAe	2289	2344	Interspersed_Repeat	c	0.7797	2.0000	76
region8	7432	7529	Mariner-N8_TAe	5	102	Interspersed_Repeat	d	0.9694	1.0000	272
region8	8104	8143	Sce_Abermu	6299	6344	Interspersed_Repeat	c	0.9286	99.0000	90
region8	9663	9802	MuDR-N111_OS	538	677	Interspersed_Repeat	d	0.7429	1.4762	142
region8	11206	11243	DNA9-31B_OS	129	173	Interspersed_Repeat	c	0.8605	1.0000	68
region8	12189	12242	Helitron-10_MT	2574	2629	Interspersed_Repeat	d	0.8214	5.0000	76
region8	12527	12635	Helitron-N12C_OS	400	499	Interspersed_Repeat	d	0.8020	4.0000	88
region8	12793	12877	HARB-1_SHS	2792	2889	Interspersed_Repeat	d	0.7978	3.0000	100
region8	12905	12985	Rep-1_AT	35	115	Interspersed_Repeat	d	0.7407	1.9091	96
region8	13163	13318	MuDR-N118_OS	519	674	Interspersed_Repeat	c	0.7516	4.1111	192

Figure 15: output3

Considering the Viridiplantae database, more repeat elements were identified, including those from other plant species. The similarity scores are also high, indicating significant matches between the repeat elements and the genomic sequence.

The summary of the different classes of repeat elements identified in the genomic region:

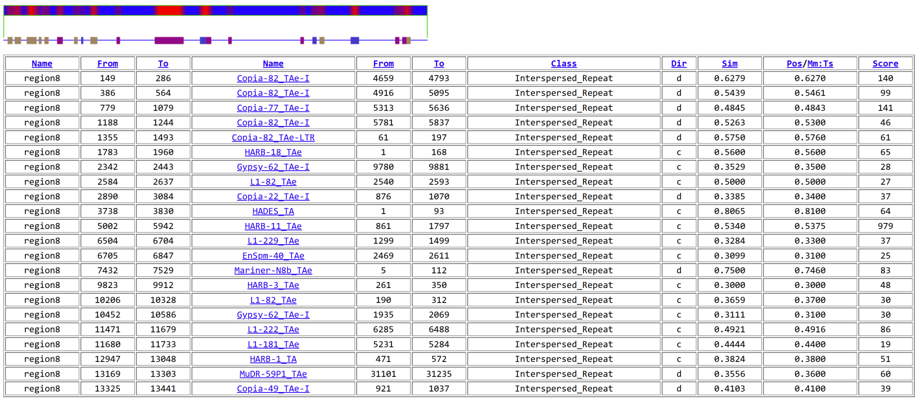
```
Repeat Class                                                    Fragments      Length
Interspersed Repeat                                                     1          81
   DNA transposon                                                      14        2117
      Harbinger                                                         3        1221
      Helitron                                                          2         163
      Mariner/Tc1                                                       3         228
      MuDR                                                              4         414
      hAT                                                               1          53
   LTR Retrotransposon                                                  5        1601
      Copia                                                             4        1527
      Gypsy                                                             1          74
   Non-LTR Retrotransposon                                              1          59
      L1                                                                1          59
Transposable Element                                                   20        3777
Total                                                                  21        3858
```

Figure 16: output4

3. **Censor with Triticum sequences and forcing translated search**:

The output can be found here: Censor Triticum sequences and translated search output.

SVG Plot and table output:



Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm:Ts	Score
region8	149	286	Copia-82_TAe-I	4659	4793	Interspersed_Repeat	d	0.6279	0.6270	140
region8	386	564	Copia-82_TAe-I	4916	5095	Interspersed_Repeat	d	0.5439	0.5461	99
region8	779	1079	Copia-77_TAe-I	5313	5636	Interspersed_Repeat	d	0.4845	0.4843	141
region8	1188	1244	Copia-82_TAe-I	5781	5837	Interspersed_Repeat	d	0.5263	0.5300	46
region8	1355	1493	Copia-82_TAe-LTR	61	197	Interspersed_Repeat	d	0.5750	0.5760	61
region8	1783	1960	HARB-18_TAe	1	168	Interspersed_Repeat	c	0.5600	0.5600	65
region8	2342	2443	Gypsy-62_TAe-I	9780	9881	Interspersed_Repeat	c	0.3529	0.3500	28
region8	2584	2637	L1-82_TAe	2540	2593	Interspersed_Repeat	c	0.5000	0.5000	27
region8	2890	3084	Copia-22_TAe-I	876	1070	Interspersed_Repeat	d	0.3385	0.3400	37
region8	3738	3830	HADES_TA	1	93	Interspersed_Repeat	c	0.8065	0.8100	64
region8	5002	5942	HARB-11_TAe	861	1797	Interspersed_Repeat	c	0.5340	0.5375	979
region8	6504	6704	L1-229_TAe	1299	1499	Interspersed_Repeat	c	0.3284	0.3300	37
region8	6705	6847	EnSpm-40_TAe	2469	2611	Interspersed_Repeat	c	0.3099	0.3100	25
region8	7432	7529	Mariner-N8b_TAe	5	112	Interspersed_Repeat	d	0.7500	0.7460	83
region8	9823	9912	HARB-3_TAe	261	350	Interspersed_Repeat	c	0.3000	0.3000	48
region8	10206	10328	L1-82_TAe	190	312	Interspersed_Repeat	c	0.3659	0.3700	30
region8	10452	10586	Gypsy-62_TAe-I	1935	2069	Interspersed_Repeat	c	0.3111	0.3100	30
region8	11471	11679	L1-222_TAe	6285	6488	Interspersed_Repeat	c	0.4921	0.4916	86
region8	11680	11733	L1-181_TAe	5231	5284	Interspersed_Repeat	c	0.4444	0.4400	19
region8	12947	13048	HARB-1_TA	471	572	Interspersed_Repeat	c	0.3824	0.3800	51
region8	13169	13303	MuDR-59P1_TAe	31101	31235	Interspersed_Repeat	d	0.3556	0.3600	60
region8	13325	13441	Copia-49_TAe-I	921	1037	Interspersed_Repeat	d	0.4103	0.4100	39

Figure 17: output5

By forcing a translated search, Censor can identify additional repeat elements that may not be detected in the standard nucleotide search. Looking at the results, they are more fragmented and have relatively low similarity scores, except for the ones already identified previously.So we will not consider this output.

RepeatMasker in DNA Subway

DNA Subway Website

DNA Subway is a bioinformatics platform that provides a suite of tools for analyzing DNA sequences, including RepeatMasker for identifying transposable elements (TEs).

The output table:

Seqid	Source	Type	Length	Start	End	Score	Strand	Phase	Attributes
wheat_53611	RepeatMasker	repeat_region	1169	130	1298	.	+	.	ID=RepeatMasker0;Name=RepeatMasker0-LTR/Copia;description=IKEROS_HV-int 5998 6933 score:5236
wheat_53611	RepeatMasker	repeat_region	280	1299	1578	.	+	.	ID=RepeatMasker1;Name=RepeatMasker1-LTR/Copia;description=IKEROS_HV-LTR 1 278 score:1428

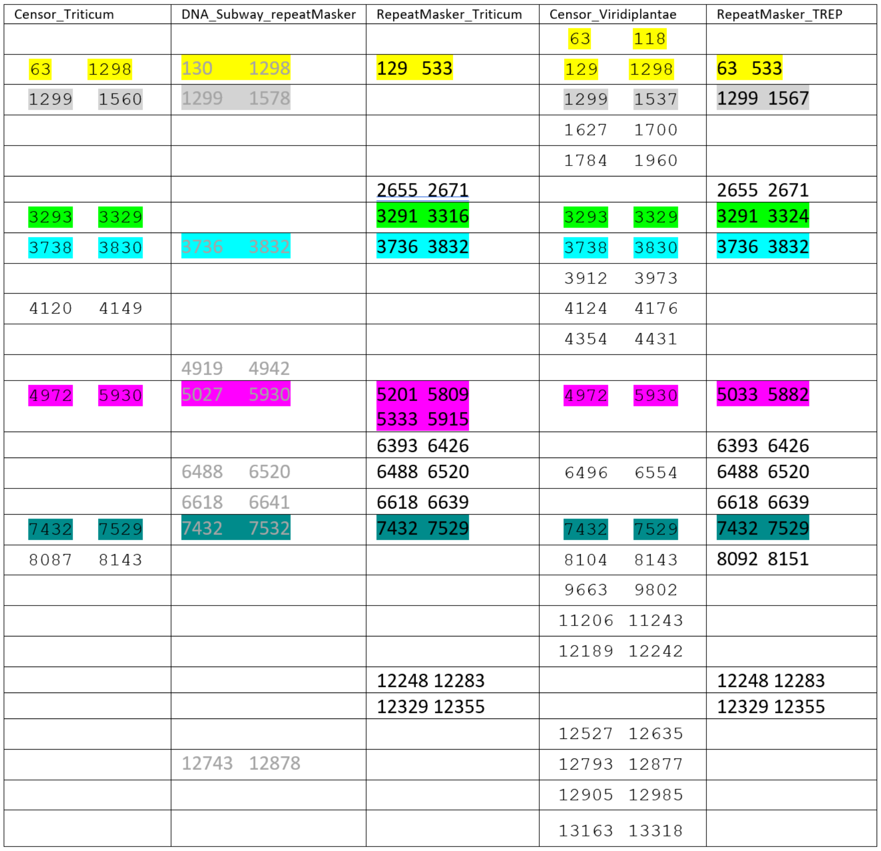
Seqid	Source	Type	Length	Start	End	Score	Strand	Phase	Attributes
wheat_53611	RepeatMasker	repeat_region	97	3736	3832	.	-	.	ID=RepeatMasker2;Name=RepeatMasker2-DNA/TcMar-Stowaway;description=HADES_TA 1 97 score:556
wheat_53611	RepeatMasker	repeat_region	24	4919	4942	.	+	.	ID=RepeatMasker3;Name=RepeatMasker3-Low_complexity;description=AT_rich 1 24 score:24
wheat_53611	RepeatMasker	repeat_region	904	5027	5930	.	-	.	ID=RepeatMasker4;Name=RepeatMasker4-DNA/PIF-Harbinger;description=HARB-N1_OS 848 2223 score:1160
wheat_53611	RepeatMasker	repeat_region	33	6488	6520	.	+	.	ID=RepeatMasker5;Name=RepeatMasker5-Simple_repeat;description=(TCGGG)n 1 33 score:189
wheat_53611	RepeatMasker	repeat_region	24	6618	6641	.	+	.	ID=RepeatMasker6;Name=RepeatMasker6-Low_complexity;description=GC_rich 1 24 score:24
wheat_53611	RepeatMasker	repeat_region	101	7432	7532	.	+	.	ID=RepeatMasker7;Name=RepeatMasker7-DNA/TcMar-Stowaway;description=ICARUS_TM 7 107 score:652
wheat_53611	RepeatMasker	repeat_region	136	12743	12878	.	+	.	ID=RepeatMasker8;Name=RepeatMasker8-Simple_repeat;description=(CTG)n 3 132 score:258

Interpretation

A nice way to visualize the results and the difference between the tools

It contains the positions of the repetitive elements found by each tool in ascending order. We colored similar positions identified by different tools with the same color.

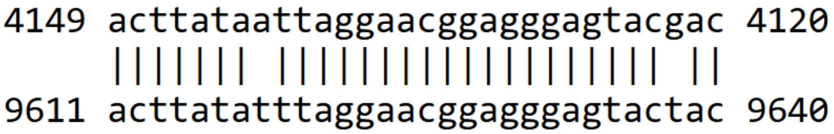
- Looking at the yellow cells, the start position 63 was agreed upon by Censor_triticum, Censor_Viridiplantae, RepeatMasker_TREP. The end position 1298 was agreed upon by Censor_triticum, Censor_Viridiplantae, and DNA Subway RepeatMasker. RepeatMasker_Triticum reported a TE inside this range but with a different start and end position, and noting that Censor_Viridiplantae reported same start and end but 2 fragments. While we cannot be sure of the exact positions, all the tools have agreed that there is a TE present in this region. If we look at the alignment provided by censor_triticum, it is globally relatively well aligned with the with `Copia-82_TAe-I`, annotated as LTR retrotransposon found in common wheat.



Censor_Triticum	DNA_Subway_repeatMasker	RepeatMasker_Triticum	Censor_Viridiplantae	RepeatMasker_TREP
			63 118	
63 1298	130 1298	129 533	129 1298	63 533
1299 1560	1299 1578		1299 1537	1299 1567
			1627 1700	
			1784 1960	
		2655 2671		2655 2671
3293 3329		3291 3316	3293 3329	3291 3324
3738 3830	3736 3832	3736 3832	3738 3830	3736 3832
			3912 3973	
4120 4149			4124 4176	
			4354 4431	
	4919 4942			
4972 5930	5027 5930	5201 5809 5333 5915	4972 5930	5033 5882
		6393 6426		6393 6426
	6488 6520	6488 6520	6496 6554	6488 6520
	6618 6641	6618 6639		6618 6639
7432 7529	7432 7532	7432 7529	7432 7529	7432 7529
8087 8143			8104 8143	8092 8151
			9663 9802	
			11206 11243	
			12189 12242	
		12248 12283		12248 12283
		12329 12355		12329 12355
			12527 12635	
	12743 12878		12793 12877	
			12905 12985	
			13163 13318	

Figure 18: summary table

- The grey cells represent the positions where Censor_Triticum, DNA Subway RepeatMasker, Censor_Viridiplantae, and RepeatMasker_TREP agreed on the start position 1299 and end position with few nucleotides differences. The TE found by Censor_Triticum in this region is also `Copia-82_TAe-LTR`.
- The green cells represent TE agreed upon by Censor and RepeatMasker. The sequence aligned to with 86% similarity using Censor is `Mariner-N15_TAe` annotated as DNA transposon, non autonomous, from Triticum Aestivum.
- The blue cells represent a TE where all the tools agreed on. While Censor gave it in the negative strand, RepeatMasker gave it in the positive strand, but DNA Subway gave it also in the negative strand. The sequence aligned there identified by the different tools is `HADES_TA`, also annotated by Censor_Triticum as Triticum aestivum non-autonomous DNA transposon.
- The pink cells also represent a TE where all of them have identified. In censor it was aligned to `HARB-11_TAe` annotated as DNA transposon from Triticum Aestivum. In RepeatMasker Triticum, there are two TEs identified around this position, the second correspond to triticum monococcum and the first to triticum aestivum `DTH_Taes_Rong_AY951945-1` which is also DNA transposon.
- For the last color, they also all agreed on it, with four of them identifying exactly same start and end positions, and they all identified it in the positive strand. It is also identified by both Censor `Mariner-N8_TAe` and RepeatMasker `DTT_Taes_Icarus_BJ274200` as DNA Transposon, Mariner superfamily. It is interesting to note that the first gene identified is from position 6226 to 10861 with intron positions: 6763-8713, CDS1: 6532- 6762, CDS2: 8714-10606. Thus, the TE is found in the intron region of the gene.
- The TE at position 4120 to 4149 identified by Censor_Triticum is with 93% similarity to `EnSpm-5n_TAe` annotated as DNA transposon from Triticum Aestivum. This TE was not identified by RepeatMasker. It is worth noting that the alignment is small:

  ```
  4149 acttataattaggaacggagggagtacgac 4120
       ||||||| ||||||||||||||||||| ||
  9611 acttatatttaggaacggagggagtactac 9640
  ```

  but still might be considered as a valid TE.
- The TE at position 8087 to 8143 identified by Censor_Triticum is with 81% similarity to `L1-169_TAe` annotated as Non-LTR retrotransposon from common wheat. Close positions were also identified by Censor_Viridiplantae and RepeatMasker_TREP.

- As for the other positions identified by RepeatMasker, they correspond mostly to simple repeats and low complexity regions. And for the other ones identified by Censor_Viridiplantae, they correspond to repeats from other species, with many intersecting with the positions of CDS. (Gene 2 positions:12512-13440, CDS1: 12512-12983, CDS2: 13169-13440).

So, in conclusion, we can rely on the TEs identified by Censor using only Triticum sequences from the Repbase as they are well annotated and mostly agreed upon by the other tools, and we do not want to include other TEs from other species that we are not sure of.

Final annotation

Gene 1 of Augustus (predicted protein) which has a length of 707aa has shown to perfectly align with a subject of the protein **"Anaphase-promoting complex subunit 11"** which also has the same length, so this is our first final annotated gene, with positions as reported by AUGUSTUS:

Feature	Start	End
gene	6226	10861
transcript	6226	10861
exon	6226	6762
start_codon	6532	6534
initial	6532	6762
terminal	8714	10606
intron	6763	8713
CDS	6532	6762
CDS	8714	10606
exon	8714	10861
stop_codon	10604	10606
tts	10861	10861

What's Anaphase-promoting complex subunit 11?

keywords: *Metal-binding, Zinc, Zing-finger, Anaphase-promoting complex subunit 11, RING*
This is an unreviewed protein annotation (TrEMBL) with score 1/5, no structure has been experimentally determined which weaken its annotation status.

It has a RING type and VWFA domains, 2 exons protein, and it has an 670 aa isoform (which explains hits of this length and possibly FGENESH's shorter prediction)

Gene 2 that we conclude is from FGENESH's gene4: **"Uncharacterized protein"** with a length of 247aa, has shown to align with a subject of the protein **"Uncharacterized protein"** which also has the same length, so this is

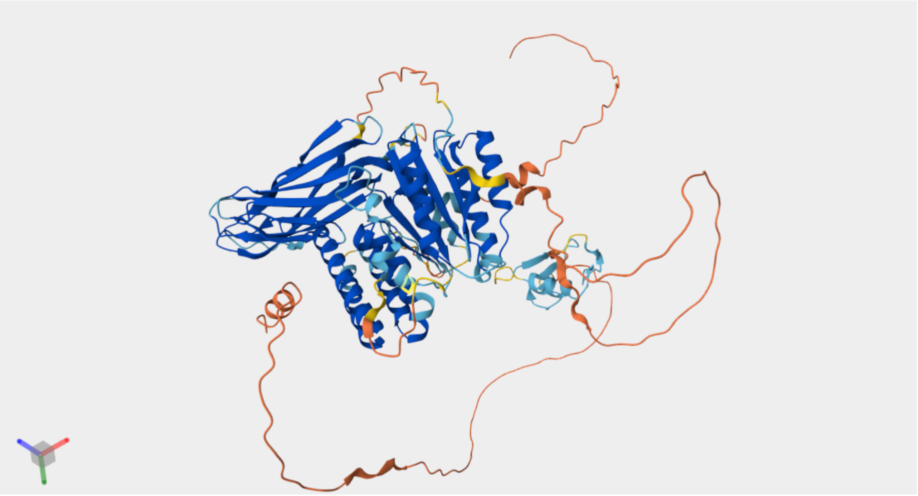
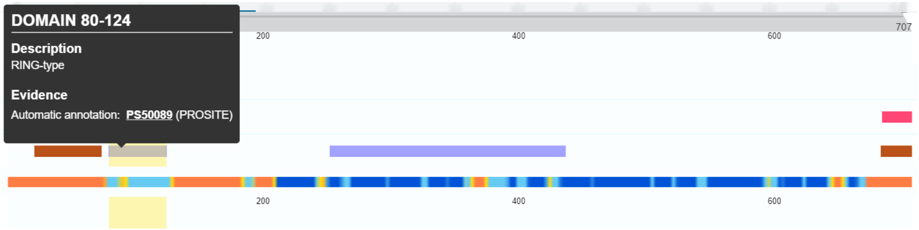


Figure 19: Anaphase-promoting complex subunit 11 predicted structure from AlphaFold



DOMAIN 80-124

Description
RING-type

Evidence
Automatic annotation: PS50089 (PROSITE)

Figure 20: Ring type domain which matches with the A. tauschii previously discussed hit

our second final annotated gene, AUGUSTUS matched teh last 245 aa of this protein. Positions as per FGENESH:

G Str	Feature	Start	End
-	PolA	11635	
-	1 CDSl	12512	12983
-	2 CDSf	13169	13440
-	TSS	13593	

What's this ambiguous "*Uncharacterized protein*"?

Getting back to this entry from it uniprot id, we notice it's involved in transcription regulation, DNA binding, and it's localized in the nucleus, it's poorly annotated as it is not reviewed (TrEMBL) and has a score of 1/5, no structure has been experimentally determined too.

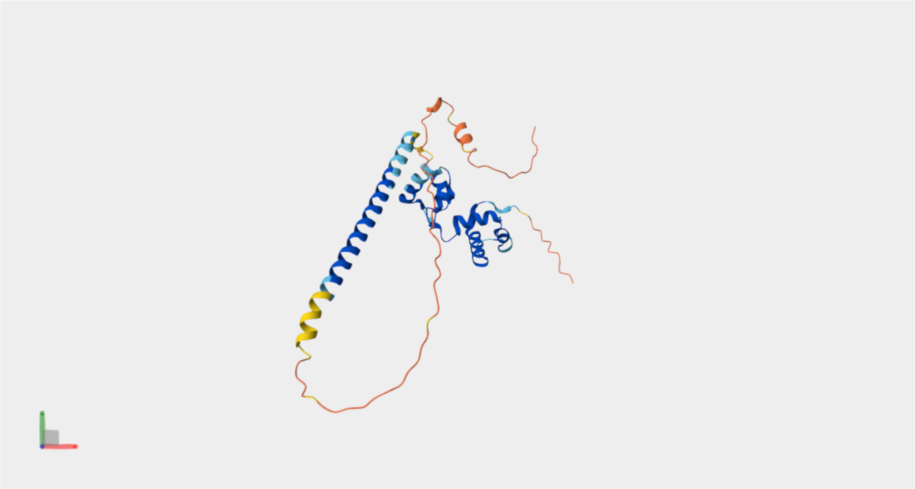


Figure 21: Uncharacterized protein predicted structure from AlphaFold

The final genes in `.fasta`:

```
>gene1   2 exons  6226  -   10881   707 aa, chain +
MADAWGRAKRALATKLCIRLPDRQRALEDAPPPPPPGREAHHPTTAVEAGPATGEEKARS
PSVSSRRLSSSGSRGSKRVCAICLGSMRTGHGQALFTAECSHKFHFHCITSNVRHGNHIC
PICRADWKELPFQGPQLADATHGRARVSPVNWPQDDGHMAVIRRLSNSYSGNLLEQFPVF
RTPEADIFNDDEQIDIQSETVEDSNAVTGSVEIKTYAEVQAIQQSVTQKVFSILIHLKAP
KSLESVSSRAPLDLVTVLDVSGSMKGAKLALLKKAMGFVIQTLGPNDRLSVIAFSSTARR
LFPLRQMNVNGRMQAMHAVNSLVDGGGTNISDGLKKGAKVIEHRRLKNPVCSIILLSDGQ
DTYSVPTFDDGVQTNHSMLVPPSILPGTGNHVQIHTFGFGADHDSAAMHAIAETSSGTFS
FIDAEGSIQNGFAQCIGGLLSVVVKEMRLGVECVDEGVVLTSIKSGGYASEVAVDGRNGS
```

```
VDIGDLYADEERGFLITLHVPAAQGQQTVLIKPSCTYQDAVTTESIQVHGSEVSVERPAY
SVDCKMSPEVEREWHRVQAMEDMSAARAAADGGDFSQAVSILEGRTRILESQAAQSSDSQ
CLALITELREMQERVESRRRYDESGRAFMLAGLSSHSWQRATARGDSTELNTQIHTYQTP
SMVDMLHRSQTLVPAVVEMLNRSPTVAPSRGSGRSVRSTKSFSERLA
>gene2   2 exons   12512   -   13440   247 aa, chain -
MAMDAMSSAVLQGAWRKGPWTALEDRLLTEYVQQQGEGSWNSVAKLTGLRRSGKSCRLRW
VNYLRPDLKRGKITADEETVILQLHAMLGNRWSAIARCLPGRTDNEIKNYWRTHFKKARP
SRRARAQLLHQYQLQQQQQHRQYLHALHLLQQQQQEMQMQLQMEQQTHQPQVMMMQQQSP
PEEDQAVITTVGNMNSMEAAECYCPCPAASAVLDLPLPADDEDALWDSLWRLVDGEDGSS
GGDSGEY
```

in .gff3:

```
##gff-version 3 format
region8 AUGUSTUS    gene    6226    10861   0.03    +   .   ID=gene1
region8 AUGUSTUS    mRNA    6226    10861   0.03    +   .   ID=gene1.t1;Parent=gene1
region8 AUGUSTUS    exon    6226    6762    .   +   .   ID=gene1.exon1;Parent=gene1.t1
region8 AUGUSTUS    CDS 6532    6762    0.94    +   0   ID=gene1.cds1;Parent=gene1.t1
region8 AUGUSTUS    intron  6763    8713    .   +   .   ID=gene1.intron1;Parent=gene1.t1
region8 AUGUSTUS    exon    8714    10861   .   +   .   ID=gene1.exon2;Parent=gene1.t1
region8 AUGUSTUS    CDS 8714    10606   0.93    +   0   ID=gene1.cds2;Parent=gene1.t1
region8 FGENESH gene    12512   13440   .   -   .   ID=gene2
region8 FGENESH mRNA    12512   13440   .   -   .   ID=gene2.t1;Parent=gene2
region8 FGENESH exon    12512   12983   .   -   .   ID=gene2.exon1;Parent=gene2.t1
region8 FGENESH CDS 12512   12983   182.93  -   0   ID=gene2.cds1;Parent=gene2.t1
region8 FGENESH exon    13169   13440   .   -   .   ID=gene2.exon2;Parent=gene2.t1
region8 FGENESH CDS 13169   13440   116.90  -   0   ID=gene2.cds2;Parent=gene2.t1
```

Running viz on artemis:

```
# we installed artemis as mentioned in file utils/tools_installation.sh on github
tools/artemis/art data/region8
# then add output/gene_final_annotation.gff
# all these files can be found on the github repo
```

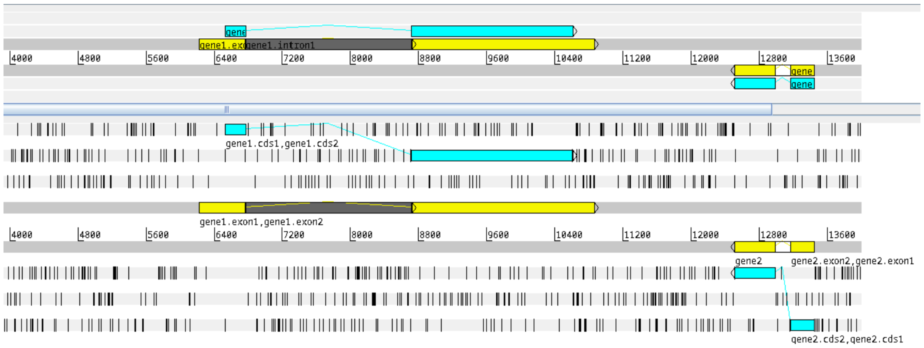


Figure 22: Final annotated and validated genes on artemis

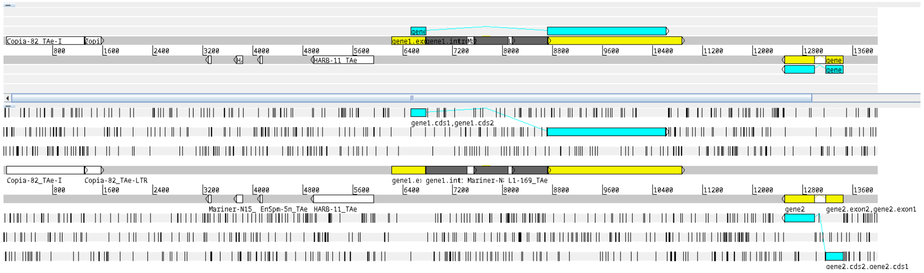


Figure 23: Final annotation of genes and TEs in white

Supplementary

- Whole Genome (all 7n chr) of *triticum aestivum* on ENSEMBL : https://ftp.ensemblgenomes.ebi.ac.uk/pub/plants/release-60/gff3/triticum_aestivum/

- ENSEMBL in general : https://plants.ensembl.org/Triticum_aestivum/Info/Index

- ENA: https://www.ebi.ac.uk/ena/browser/view/Taxon:4565
- SRA: Sequence Read Archive, repository for seq data

- RNAseq reads fetch and viz: youtube video

- RefSeq: reference sequence v2.1 here, link to acces the dataset is *here*

- downloading a proteome of a species from uniprot, EMBL-EBI training course

- Chromosome 4D annotations in GFF *ftp link*