université PARIS-SACLAY

STRUCTURAL GENOMICS

Structural Annotations of Region 7 of Wheat's genome

Authors:

Talissa KASSABLY / 20222483

Bruno YOUNG DE CASTRO / 20251333

Akshay MOHAMMED KHOKAN / 20224297

M1 GENIOMHE 2025-2026

Table of contents

Structural Annotations of Region 7 of Wheat's genome.

Bruno YOUNG DE CASTRO, Talissa KASSABLY & Akshay MOHAMMED KHOKAN

INTRODUCTION

The Wheat (*Triticum aestivum*) genome is one that is widely known for its complexity and the difficulties aligned with its annotation. It was found by the researchers involved in the project by the International Wheat Genome Sequencing Consortium that the wheat genome was 5 times larger than the human genome and 85% of the genome is made up of repetitive sequences. This genome is hexaploidy, meaning that there is a total of 6 copies of each chromosome, with three closely related sub genomes (A, B and D). With this project they were able to find a total of 107,891 genes and 4.7 million molecular markers.[1]

In this project we have been tasked to perform the gene annotation of approximately a 16kB region of the wheat genome, where we are tasked to identify genes (complete coordinates of intron-exon structure and UTR, validation by the presence of transcribed sequences and/or homologous genes), proteins (potential protein functions, motifs and domains), and transposable elements (coordinates and family of the transposable element). The gene annotation of the wheat genome was performed through a dedicated pipeline that we developed. The pipeline was designed as a notebook, mimicking the steps that would be performed in a laboratory. All files used for this project, including the notebook and the results, are provided in the following GitHub repository:

https://github.com/crakshay1/BruAnnoPipe.

We must be conscious of the main difficulties that may arise in our aim to annotate this genome. The hexaploidy structure with the presence of three similar sub genomes might prove to be an issue to determine to which sub genome (A, B or D) our region belongs to. It's highly repetitive nature (85% of the whole genome) can interrupt or mimic gene structures and retrotransposons can make it difficult for gene prediction software to determine the presence of actual genes with accuracy, as it might detect these structures as genes. Also, there is uncertainty about the presence of RNA-seq or cDNA data for our specific 16kB region, which will lead to predictions being made based on homology and this can prove to be a mistake in

[1] Australian Center for International Agricultural Research, 2019

such a complex genome. Finally, plant genomes commonly introduce alternative splicing, which can make it hard to determine the intron-exon boundaries

STRUCTURAL ANNOTATIONS

Genes (Pre-Masking)

For our gene predictions we will be using the Augustus tool software, we prefer this tool over others such as FGENESH because it is more conservative in gene prediction, this is because FGENESH works best on large eukaryotic genomes, and we're only analyzing a 15kB region of an eukaryotic genome, therefore, we assessed that the conservative approach of Augustus will be better, that way we reduce the number of false positives. In our case, the Augustus tool will predict the presence of genes ab initio, but it is also able to integrate external evidence such as RNA-Seq, ESTs or proteomics. This tool can be used in a website (bio.tools website) or it can be used locally. This tool creates its predictions with the use of a Hidden Markov Model (HMM) which is trained on known genes of a given organism, which allows it to predict genes and biological features such as introns, exons, start codon, stop codon and splice sites through characteristic sequence patterns. For this project, our input for the Augustus tool will be our genomic DNA sequence (in FASTA format) and we will specify the species (*Triticum aestivum*) for the tool to compare, it will output the predicted gene models, coding and protein sequences in the common General Feature Format (GFF) file[2].

Gene 1:

```
# ----- prediction on sequence number 1 (length = 15001, name = region7) -----
#
# Predicted genes for sequence number 1 on both strands
# start gene g1
region7 AUGUSTUS    gene    1305    3965    0.09    +   .   g1
region7 AUGUSTUS    transcript  1305    3965    0.09    +   .   g1.t1
region7 AUGUSTUS    exon    1305    1346    .   +   .   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    start_codon 1329    1331    .   +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    initial 1329    1346    0.68    +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    terminal    1487    3676    0.87    +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    intron  1347    1486    0.85    +   .   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    CDS 1329    1346    0.68    +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    CDS 1487    3676    0.87    +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    exon    1487    3965    .   +   .   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    stop_codon  3674    3676    .   +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    tts 3965    3965    .   +   .   transcript_id "g1.t1"; gene_id "g1";
```

[2] Stanke M., et al. (2006)

This first predicted gene is shown to have 2 exons and 1 intron; we can also observe that it is present on the forward strand based on the "+" symbol and that the whole gene goes from position 1305 to 3965. On this sequence our prediction finds the first exon to start at position 1305 and has the start codon in position 1329-1331. The end of this first exon is at position 1346, where the intron starts from position 1347 until position 1486. The second exon begins at position 1487 and ends at position 3965, however, the stop codon is found at position 3674-3676. The first coding sequence is found to cover from the start codon (1329) until the end of the first exon (1346) and the second coding sequence covers from the start of the second exon (1487) until the end of the stop codon (3676).

Coding Sequence:

[atggaagcccctgaccaggagaatccttgcgccatctgcctcggcggcatggccgccggcggcgggcaggccacc
ttcacggcggagtgctcccacaccttccacttcaactgcatctccgccagcgtcgcgcacggccacctcgtctgccc
gctctgcaacgcgcgctggcgagagctgcccttcctgcgacccaccgcgccggtgccgcagccgcctacgctgcct
aggctgggtcgtcccgttcccatgcacggcgtgcagcctccgagcgagccgacagcatcgcctcctctcatgcatgg
cgggatgcctccgttcccagcgcaggcgccaccgccgccgcgcgggcatatcatgcagcatcaccagccgccgcc
gccgaacgtgcatgtcgtgcagcatcatcagccgccgccgcccgtgcataccgtgcagcatcatcagccccgccg
cccgagcctacggtcgtcttcgacgacgacgagcaggtggagccggcctccaggccgccagctgacagcacaccg
gcagctgcatcgaacggggcagtggtcgtcaacacgcacgccgagtactcggccgtcgccagggactcgtccagcg
acaacttcgccgtgctcgtgcacgtcaaggctcccgcgatggccgacaccgtggcggccggcagcgacaagccgc
ccccgcgcgcgccgctggacctcgtgaccgtgctcgacgtcagcggcagcatgagcggccacaaactggcgctcc
tgaagcaggccatgcggttcgtcatcgacaacctcggccccaacgaccgcctctccgtcgtgtccttctcctccgag
gcgcgccggctgaccaggctcacgcgcatgtcggacgccgggaaggcactggccgtgagcgccgtggagtccctc
gcggcgcgcggcggcaccaacatcgccgaggggctccgcacggccgccaaggtgctcgacgagcgccggcacag
gaacgccgtctccagcgtcgtgctcctctccgacggtcaggatacctataccatgatgaggcgccggggaccgtccg
gcgtccaggccaacaactacgaggagctcgtcccgccctccttcgcacgcacgggcgctgacggcgagtggtccg
cgccgatccacaccttcggcttcgggaacgaccacgacgcggccgcgatgcacgtcatcgccgaggcgacgggcg
gcacgttctcgttcatcgagaacgaggctgtgatacaggacgcgttcgcgcagtgcatcggcggcctgctctccgtcgt
ggtccaggaggcgcgcatcgccgtcgcgtgcgtgcaccccggggtccgtgtcgtctccgtcaagtccggccgttacg
agagccgcgtcgacgaggacggctgcgccgcatctgtccgagtcggggagctctacgccgacgaggagaggcgttt
cttgctctttctgaccgtgccaagagtcgaagcgacggacggcgacaccactgctcttgcgagagtggtcttcagcta
cagaaacgcggcgagcggcgcggaggtgagcgtgacggccgaggacacggtggtggcgaggccggagcacgcgc
cgagcgcgtcggagcgctcagtggaggtggagcgggagcgcgtccgggtggaggcggcagaggacatcgcggcgg
cgagggcagcggcggagcggggcgagcaccaggaagcggtggagatcctcgacaaccgtcagcgggcgctggag
cagtcggaggcggcaggggacggcgaccccatgatcgtggcgctgggggcggagctgcaggagatgcgcgggcgc
gtgtcgaaccggcagagctacatgcggtcggggcgggcgtacatgctggccggcatgagcgcgcaccagcagcaa
cgcgccacctccaggcagatgctggagccggaggagcagcagacgtcgatgatggcgaggaatagtggagtgagga
ggatgatcagaagaggagtggggtcgagcggcgggggatatatggcggcagcggcgcccgtggccgaggcgtcgaa

cgaggcgacgatgtcgtacgcgacgccggccatgcgcgccatgctgctgcgctcgcgggaggcgcgtggggcgtcg
gccgagcaagggcagcaggaggagcagcagcccatggccggaaaagacgatgccgggagctcgggcccgaagg
acgtgaaccaatag]

Protein Sequence:

[MEAPDQENPCAICLGGMAAGGGQATFTAECSHTFHFNCISASVAHGHLVCPLCNARWREL
PFLRPTAPVPQPPTLPRLGRPVPMHGVQPPSEPTASPPLMHGGMPPFPAQAPPPPRGHIMQ
HHQPPPPNVHVVQHHQPPPPVHTVQHHQPPPPEPTVVFDDDEQVEPASRPPADSTPAAAS
NGAVVVNTHAEYSAVARDSSSDNFAVLVHVKAPAMADTVAAGSDKPPPRAPLDLVTVLDVSG
SMSGHKLALLKQAMRFVIDNLGPNDRLSVVSFSSEARRLTRLTRMSDAGKALAVSAVESLAAR
GGTNIAEGLRTAAKVLDERRHRNAVSSVVLLSDGQDTYTMMRRRGPSGVQANNYEELVPPS
FARTGADGEWSAPIHTFGFGNDHDAAAMHVIAEATGGTFSFIENEAVIQDAFAQCIGGLLSVV
VQEARIAVACVHPGVRVVSVKSGRYESRVDEDGCAASVRVGELYADEERRFLLFLTVPRVEAT
DGDTTALARVVFSYRNAASGAEVSVTAEDTVVARPEHAPSASERSVEVERERVRVEAAEDIAA
ARAAAERGEHQEAVEILDNRQRALEQSEAAGDGDPMIVALGAELQEMRGRVSNRQSYMRS
GRAYMLAGMSAHQQQRATSRQMLEPEEQQTSMMARNSGVRRMIRRGVGSSGGGYMAAA
APVAEASNEATMSYATPAMRAMLLRSREARGASAEQGQQEEQQPMAGKDDAGSSGPKDVN
Q]

Gene 2:

```
region7 AUGUSTUS    gene    4366    5770    0.16    +    .    g2
region7 AUGUSTUS    transcript  4366    5770    0.16    +    .    g2.t1
region7 AUGUSTUS    tss 4366    4366    .    +    .    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    exon    4366    4639    .    +    .    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    start_codon 4473    4475    .    +    0    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    initial 4473    4639    0.88    +    0    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    terminal    5240    5666    0.81    +    1    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    intron  4640    5239    0.76    +    .    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    CDS 4473    4639    0.88    +    0    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    CDS 5240    5666    0.81    +    1    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    exon    5240    5770    .    +    .    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    stop_codon  5664    5666    .    +    0    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    tts 5770    5770    .    +    .    transcript_id "g2.t1"; gene_id "g2";
```

This second gene has been found to have 2 exons and 1 intron and also to be located in the forward strand as did the first one. This gene is found to start in position 4366 and end in 5770. The first exon is found to start at position 4366, same place where the transcription start site is located, however, the start codon is located downstream at position 4473-4475. The exon ends at position 4639, so the intron begins at position 4640 and ends at position 5239. The second exon, therefore, starts at position 5240 and ends in position 5770. However, the stop codon is found to be in position 5664-5666. The coding sequence covers from the start codon (4473) until

the end of the first exon (4639), and from the beginning of the second exon (5240) until the stop codon (5666).

Coding Sequence:

[atgcggggcgaccagctgcgcggcgaggtgccgggtgctgcaaccgcctatgcgtggagctggaaccggcacgcg
gccgtgcagcgaccagcacatagcggtactgctaccgacgcggcgtggcgctacgaccagcgcggcggcgacggc
caccgtcggccaaaaagcgtactccggagccgacgggggcggagacgacgtgcaggcatcgaggcgagatggaa
gcagggctactcgggagcgaaggcgatgagccgacgggggcggagcccacgtgcgggcaccaaggtgagatgga
agcggggctgctcgggtctcgggagcgaaggcagcaagccggcgggggggcgaagacgacgtgcgggcaccgag
gcaagatggaagtggggctgctcgggagggaaggcggcgagccggcaaggggcggagatgcggcgagcggtggcg
agcaaggaaggggcggagatccggcgagcttcgacgccggtcatgtcgctcgtccatggcagttctcgttcctgctct
cccaatttcagattcacgggcaaaagcgaaaaagaaacagtacaacggggacattggatcagatctga]

Protein Sequence:

[MRGDQLRGEVPGAATAYAWSWNRHAAVQRPAHSGTATDAAWRYDQRGGDGHRRPKSVL
RSRRGRRRRAGIEARWKQGYSGAKAMSRRGAEPTCGHQGEMEAGLLGSRERRQQAGGGA
KTTCGHRGKMEVGLLGREGGEPARGGDAASGGEQGRGGDPASFDAGHVARPWQFSFLLS
QFQIHGQKRKRNSTTGTLDQI]

Gene 3:

```
region7 AUGUSTUS    gene    6938    7765    0.11    -   .   g3
region7 AUGUSTUS    transcript  6938    7765    0.11    -   .   g3.t1
region7 AUGUSTUS    tts 6938    6938    .   -   .   transcript_id "g3.t1"; gene_id "g3";
region7 AUGUSTUS    exon    6938    7765    .   -   .   transcript_id "g3.t1"; gene_id "g3";
region7 AUGUSTUS    stop_codon  7155    7157    .   -   0   transcript_id "g3.t1"; gene_id "g3";
region7 AUGUSTUS    single  7155    7421    1   -   0   transcript_id "g3.t1"; gene_id "g3";
region7 AUGUSTUS    CDS 7155    7421    1   -   0   transcript_id "g3.t1"; gene_id "g3";
region7 AUGUSTUS    start_codon 7419    7421    .   -   0   transcript_id "g3.t1"; gene_id "g3";
region7 AUGUSTUS    tss 7765    7765    .   -   .   transcript_id "g3.t1"; gene_id "g3";
```

This gene is found to be in the reverse strand and goes from position 6938 until position 7765 and is composed of a single exon. The start codon is found in position 7419 and the stop codon in position 7155. The coding sequence in this gene covers from the start codon (7421) until the stop codon (7155). Recall that the start codon is located downstream the gene in comparison to the stop codon because we are in the reverse strand, therefore, transcription will be performed in the opposite direction than in the forward strand.

Coding Sequence:

[atgggctactttccagcgcctcgaggggatcttcgatccttccatccagcgcctctgggcggccgaggacgcatgctg
gaggcgaaggaagccgccgctgccagagatgcagctgcggctgcgcacctgccggcgctccgggaggagtaccaa
gtcgcttcgaccctctccgacacgctcatctaccataccctccacggcaacgagcccccaccctgcaacggcgac
tcggacgacaacaacgactcggcgaggagtagttgctaa]

Protein Sequence:

[MGYFPAPRGDLRSFHPAPLGGRGRMLEAKEAAAARDAAAAAHLPALREEYQVASTLSDTLIY
HTLHGNEPPPCNGDSDDNNDSARSSC]

Gene 4:

```
region7 AUGUSTUS    gene    11536   12971   0.02    +   .   g4
region7 AUGUSTUS    transcript  11536   12971   0.02    +   .   g4.t1
region7 AUGUSTUS    tss 11536   11536   .   +   .   transcript_id "g4.t1"; gene_id "g4";
region7 AUGUSTUS    exon    11536   11715   .   +   .   transcript_id "g4.t1"; gene_id "g4";
region7 AUGUSTUS    start_codon 11699   11701   .   +   0   transcript_id "g4.t1"; gene_id "g4";
region7 AUGUSTUS    initial 11699   11715   0.53    +   0   transcript_id "g4.t1"; gene_id "g4";
region7 AUGUSTUS    terminal    11928   12486   0.49    +   1   transcript_id "g4.t1"; gene_id "g4";
region7 AUGUSTUS    intron  11716   11927   0.5 +   .   transcript_id "g4.t1"; gene_id "g4";
region7 AUGUSTUS    CDS 11699   11715   0.53    +   0   transcript_id "g4.t1"; gene_id "g4";
region7 AUGUSTUS    CDS 11928   12486   0.49    +   1   transcript_id "g4.t1"; gene_id "g4";
region7 AUGUSTUS    exon    11928   12971   .   +   .   transcript_id "g4.t1"; gene_id "g4";
region7 AUGUSTUS    stop_codon  12484   12486   .   +   0   transcript_id "g4.t1"; gene_id "g4";
region7 AUGUSTUS    tts 12971   12971   .   +   .   transcript_id "g4.t1"; gene_id "g4";
```

This is the final gene found in our region and is composed of two exons and in the forward strand. This gene is found between positions 11536 and position 12971. The first exon is found between positions 11536 and position 11715, where the start codon is found at position 11699. After the end of the exon, we find an intron which comprises from position 11716 until position 11927, then the second exon starts in position 11928 and is found to last until position 12971. In the second exon we find the stop codon at position 12484.

Coding Sequence :

[atggccggaacgctaagttgcgccgccgcaacggccgcgccgttcctctcgatccggacgggcatcggcaagccg
agccgttcggaagaagaactcaccccacaggggcaaagggcaccaccacagagccatttgacggcgacgcgttc
accctcgaggtgctcgcgcacgccggcaaaggggtcagcgggggcgccgcatctggaaaccgccgtcgctcgcgc
gaaggggagcagaggaggctctgctcgcgctgctcgcgtccctctcgcgctggggtggtggtggaagaagagaggga
aggggaacggctcgcgcggcgcagtggccctcgccgccgagcacgggggccggagcggcctaccagggcgctgc
ccttttaccttgacagagagggagccgacgggacagaagaagagaaagaagaagggaaaggggggaagaaagag
aacggccaggctcgcgcggtggtggacgccgtcgccggcggctcgaccggacaggcacggcgcggcggccgagc
acggcgcggtcgccggcgaccgggacatcgtcgggcgagagacgagaggatga]

Protein Sequence :

[MAGTLSCAAATAAPFLSIRTGIGKPSRSEEELTPQGQRAPPQSHLTATRSPSRCSRTPAKGSA
GAPHLETAVARAKGSRGGSARAARVPLALGWWWKKRGKGNGSRGAVALAAEHGGRSGLP
GRCPFYLDREGADGTEEEKEEGKGGRKRTARLARWWTPSPAARPDRHGAAAEHGAVAGDR
DIVGRETRG]

Transposable Elements

Transposable elements are mobile DNA sequences capable of replicating themselves within genomes, and specifically in the wheat genome, where 85% of its genome is found to be transposable elements as previously explained. These sequences can severely disrupt the gene prediction process, as it can lead to a higher number of genes than there actually are, which is why repeat masking is recommended before gene prediction is performed[3].

To detect transposable elements and obtain a masked sequence of our region, we will be using the Censor tool. We have also done this prediction with the BlastN tool against the TrepDB database, however, when we validated the results with the Dotplot tool (further explained in the validation section), we have seen that the results are not supported, which is why we have chosen to continue our analysis with the Censor tool, as its results are supported by our validation. This tool detects transposable elements and other DNA repeats by comparing our input DNA sequence and comparing it against a curated library of known repeats such as Repbase. It will take our original sequence in the FASTA format as an input and it will output our masked sequence (soft: lowercase letters/ hard: repeat regions replaced with X) and a repeat annotation report on what it found.

After the analysis with the Censor tool, we obtain the following results.

Summary Table

Repeat Class	Fragments	Length
Transposable Element	16	4718
DNA transposon	11	2248
EnSpm/CACTA	2	227
Mariner/Tc1	3	378
MuDR	6	1643
LTR Retrotransposon	3	1893
Copia	3	1893
Non-LTR Retrotransposon	2	577
L1	2	577
Total	**16**	**4718**

[3] Hoff, K. J., & Stanke, M. (2019).

Map of Hits

SVG viewer is required to view graphical representation of the map as Scalable Vector Graphics (SVG plot).

region7 (SVG Plot; Alignments; Masked)

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/ Mm:Ts	Score
region7	72	567	L1-384_TAe	1742	2250	Interspersed_Repeat	c	0.6755	2.5435	108
region7	700	802	Mariner-N8_TAe	1	103	Interspersed_Repeat	c	0.9126	1.8000	246
region7	2392	2472	L1-N6_TAe	136	205	Interspersed_Repeat	c	0.7568	4.6667	70
region7	3826	4008	MuDR-N38_TAe	1	182	Interspersed_Repeat	d	0.9667	2.0000	478
region7	5443	5557	MuDR-92_TAe	533	661	Interspersed_Repeat	c	0.7417	2.7143	74
region7	5658	5740	MuDR-84_TAe	8	90	Interspersed_Repeat	c	0.7750	2.6667	92
region7	6358	6416	MuDR-39_TAe	5188	5242	Interspersed_Repeat	d	0.8246	2.5000	76
region7	6424	6520	Mariner-N10_TAe	7	103	Interspersed_Repeat	d	0.9175	1.0000	234
region7	6651	6720	EnSpm-24c_TAe	575	652	Interspersed_Repeat	d	0.9041	4.0000	154
region7	6918	7095	Mariner-N10_TAe	39	215	Interspersed_Repeat	d	0.9326	2.2000	446
region7	8135	8576	MuDR-N69_TAe	371	805	Interspersed_Repeat	d	0.6989	3.0800	138
region7	8697	8853	EnSpm-5h3_TAe	11782	11953	Interspersed_Repeat	c	0.8428	1.9167	270
region7	9798	10558	MuDR-59N1_TAe	1	968	Interspersed_Repeat	d	0.8410	1.6415	978
region7	11389	11650	Copia-82_TAe-LTR	1	254	Interspersed_Repeat	d	0.7665	1.8000	272
region7	11844	13112	Copia-82_TAe-I	38	1286	Interspersed_Repeat	d	0.7776	2.8421	1082
region7	14634	14995	Copia-82_TAe-I	4018	4376	Interspersed_Repeat	d	0.7280	1.8837	322

The tool was able to find a total of 16 transposable elements in our sequence, comprising 4.7kB (approximately 25% of the whole sequence). 11 of them where DNA transposons, which move between genomic locations by a copy-and-paste mechanism, from the Mariner/Tc1 family (3 out of 11), MuDR family (6) and EnSpm/CACTA family (2). It is important to note that, for the hits of the Mariner/Tc1 and EnSpm/CACTA families, we only find small fragments (<200bp) at high similarity, which can note to recent insertions or conserved family motifs, that can still disrupt gene prediction. On the other hand, the high abundance and length of the fragments of the MuDR family can strongly interfere with gene prediction if not masked. 3 of the hits were LTR Retrotransposons signals of a Copia-82 element, but only one of the 3 was found to be an LTR, which is unusual as normally we will observe the internal region being flanked by two LTRs not only one. This suggests that we might only have a partial copia-82 element and not the complete structure. Finally, 2 were found to be hits of Non-LTR Retrotransposons

(L1), and even though one is very short (81bp) and the other has a low similarity (<0.7), they can still affect gene prediction and must be masked.

Gene Prediction (Post-Masking)

After the analysis of the transposable elements present, we will perform gene predictions with Augustus on the masked sequence we have obtained from the Censor tool. Augustus show that from the 4 predicted genes previously, with our masked sequence, now we only observe 2 predicted genes, this highlights the importance of transposable element analysis and how performing annotation with an unmasked sequence can lead to incorrect results.

Gene 1:

```
region7 AUGUSTUS    gene    1305    4122    0.14    +   .   g1
region7 AUGUSTUS    transcript  1305    4122    0.14    +   .   g1.t1
region7 AUGUSTUS    exon    1305    1346    .   +   .   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    start_codon 1329    1331    .   +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    initial 1329    1346    0.66    +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    terminal    1487    3676    0.84    +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    intron  1347    1486    0.76    +   .   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    CDS 1329    1346    0.66    +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    CDS 1487    3676    0.84    +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    exon    1487    4122    .   +   .   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    stop_codon  3674    3676    .   +   0   transcript_id "g1.t1"; gene_id "g1";
region7 AUGUSTUS    tts 4122    4122    .   +   .   transcript_id "g1.t1"; gene_id "g1";
```

This first gene starts at position 1305 and ends at position 4122, it presents 2 exons and 1 intron and is located on the forward strand. The first exon covers from position 1305 to position 1346 with the start codon being found in position 1329-1331. The intron covers from the end of the first exon at position 1347 until position 1486. The second exon starts at position 1487 and covers until position 4122, with the stop codon being found in position 3674-3676. The first coding sequence is found to cover from the start codon (1329) until the end of the first exon (1346) and the second coding sequence covers from the start of the second exon (1487) until the end of the stop codon (3676). This gene is very similar to gene 1 found previously to the masking, has the same positions for all the key elements (start codon, stop codon and intron-exon boundaries) but the end was found ≈150bp further downstream. However, there are 81bp that are masked, therefore we cannot know if they are identical.

Coding Sequence:

[atggaagcccctgaccaggagaatccttgcgccatctgcctcggcggcatggccgccggcggcgggcaggccacc
ttcacggcggagtgctcccacaccttccacttcaactgcatctccgccagcgtcgcgcacggccacctcgtctgccc
gctctgcaacgcgcgctggcgagagctgcccttcctgcgacccaccgcgccggtgccgcagccgcctacgctgcct
aggctgggtcgtcccgttcccatgcacggcgtgcagcctccgagcgagccgacagcatcgcctcctctcatgcatgg
cgggatgcctccgttcccagcgcaggcgccaccgccgccgcgcgggcatatcatgcagcatcaccagccgccgcc
gccgaacgtgcatgtcgtgcagcatcatcagccgccgccgcccgtgcataccgtgcagcatcatcagcccccgccg
cccgagcctacggtcgtcttcgacgacgacgagcaggtggagccggcctccaggccgccagctgacagcacaccg
gcagctgcatcgaacggggcagtggtcgtcaacacgcacgccgagtactcggccgtcgccagggactcgtccagcg
acaacttcgccgtgctcgtgcacgtcaaggctcccgcgatggccgacaccgtggcggccggcagcgacaagccgc
ccccgcgcgcgccgctggacctcgtgaccgtgctcgacgtcagcggcagcatgagcggccacaaactggcgctcc
tgaagcaggccatgcggttcgtcatcgacaacctcggccccaacgaccgcctctccgtcgtgtccttctcctccgag
gcgcgccggctgaccaggctcacgcgcatgtcggacgccgggaaggcactggccgtgagcgccgtggagtccctc
gcggcgcgcggcggxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxgctcctctccgacggtcaggatacctataccatgatgaggcgccggggaccgtccggcgtcc
aggccaacaactacgaggagctcgtcccgccctccttcgcacgcacgggcgctgacggcgagtggtccgcgccga
tccacaccttcggcttcgggaacgaccacgacgcggccgcgatgcacgtcatcgccgaggcgacgggcggcacgtt
ctcgttcatcgagaacgaggctgtgatacaggacgcgttcgcgcagtgcatcggcggcctgctctccgtcgtggtcca
ggaggcgcgcatcgccgtcgcgtgcgtgcaccccggggtccgtgtcgtctccgtcaagtccggccgttacgagagcc
gcgtcgacgaggacggctgcgccgcatctgtccgagtcggggagctctacgccgacgaggagaggcgtttcttgctct
ttctgaccgtgccaagagtcgaagcgacggacggcgacaccactgctcttgcgagagtggtcttcagctacagaaac
gcggcgagcggcgcggaggtgagcgtgacggccgaggacacggtggtggcgaggccggagcacgcgccgagcgc
gtcggagcgctcagtggaggtggagcgggagcgcgtccgggtggaggcggcagaggacatcgcggcggcgagggc
agcggcggagcggggcgagcaccaggaagcggtggagatcctcgacaaccgtcagcgggcgctggagcagtcgga
ggcggcaggggacggcgaccccatgatcgtggcgctgggggcggagctgcaggagatgcgcgggcgcgtgtcgaac
cggcagagctacatgcggtcggggcgggcgtacatgctggccggcatgagcgcgcaccagcagcaacgcgccacc
tccaggcagatgctggagccggaggagcagcagacgtcgatgatggcgaggaatagtggagtgaggaggatgatca
gaagaggagtggggtcgagcggcgggggatatatggcggcagcggcgcccgtggccgaggcgtcgaacgaggcga
cgatgtcgtacgcgacgccggccatgcgcgccatgctgctgcgctcgcgggaggcgcgtggggcgtcggccgagca
agggcagcaggaggagcagcagcccatggccggaaaagacgatgccgggagctcgggcccgaaggacgtgaacc
aatag]

Protein Sequence:

[MEAPDQENPCAICLGGMAAGGGQATFTAECSHTFHFNCISASVAHGHLVCPLCNARWREL
PFLRPTAPVPQPPTLPRLGRPVPMHGVQPPSEPTASPPLMHGGMPPFPAQAPPPPRGHIMQ
HHQPPPPNVHVVQHHQPPPPVHTVQHHQPPPPEPTVVFDDDEQVEPASRPPADSTPAAAS
NGAVVVNTHAEYSAVARDSSSDNFAVLVHVKAPAMADTVAAGSDKPPPRAPLDLVTVLDVSG
SMSGHKLALLKQAMRFVIDNLGPNDRLSVVSFSSEARRLTRLTRMSDAGKALAVSAVESLAAR
GXXXXXXXXXXXXXXXXXXXXXXXXXXXXLLSDGQDTYTMMRRRGPSGVQANNYEELVPPSFA

RTGADGEWSAPIHTFGFGNDHDAAAMHVIAEATGGTFSFIENEAVIQDAFAQCIGGLLSVVV
QEARIAVACVHPGVRVVSVKSGRYESRVDEDGCAASVRVGELYADEERRFLLFLTVPRVEATD
GDTTALARVVFSYRNAASGAEVSVTAEDTVVARPEHAPSASERSVEVERERVRVEAAEDIAAA
RAAAERGEHQEAVEILDNRQRALEQSEAAGDGDPMIVALGAELQEMRGRVSNRQSYMRSG
RAYMLAGMSAHQQQRATSRQMLEPEEQQTSMMARNSGVRRMIRRGVGSSGGGYMAAAA
PVAEASNEATMSYATPAMRAMLLRSREARGASAEQGQQEEQQPMAGKDDAGSSGPKDVN
Q]

Gene 2:

```
region7 AUGUSTUS    gene    6774    7765    0.09    -    .    g2
region7 AUGUSTUS    transcript  6774    7765    0.09    -    .    g2.t1
region7 AUGUSTUS    tts 6774    6774    .    -    .    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    exon    6774    7765    .    -    .    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    stop_codon  7155    7157    .    -    0    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    single  7155    7421    0.92    -    0    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    CDS 7155    7421    0.92    -    0    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    start_codon 7419    7421    .    -    0    transcript_id "g2.t1"; gene_id "g2";
region7 AUGUSTUS    tss 7765    7765    .    -    .    transcript_id "g2.t1"; gene_id "g2";
```

This second gene is found to cover from position 6774 to position 7765 in the reverse strand and only has one exon, which covers the whole gene. The transcription start site is found in position 7765, with the start codon in position 7419-7421 and the stop codon in position 7155-7157. The coding sequence for this gene covers from the start codon (7421) to the stop codon (7419). Recall that the start codon is located downstream the gene in comparison to the stop codon because we are in the reverse strand, therefore, transcription will be performed in the opposite direction than in the forward strand. This gene is very similar to gene 3 found previously to the masking, has the same positions for all the key elements (start codon, stop codon and transcription start site) but the end was found ≈150bp further upstream. Furthermore, they share the exact same coding and protein sequences.

Coding Sequence:

[atgggctactttccagcgcctcgaggggatcttcgatccttccatccagcgcctctgggcggccgaggacgcatgctg
gaggcgaaggaagccgccgctgccagagatgcagctgcggctgcgcacctgccggcgctccgggaggagtaccaa
gtcgcttcgaccctctccgacacgctcatctaccataccctccacggcaacgagcccccaccctgcaacggcgac
tcggacgacaacaacgactcggcgaggagtagttgctaa]

Protein Sequence:

[MGYFPAPRGDLRSFHPAPLGGRGRMLEAKEAAAARDAAAAAHLPALREEYQVASTLSDTLIY
HTLHGNEPPPCNGDSDDNNDSARSSC]

VALIDATIONS

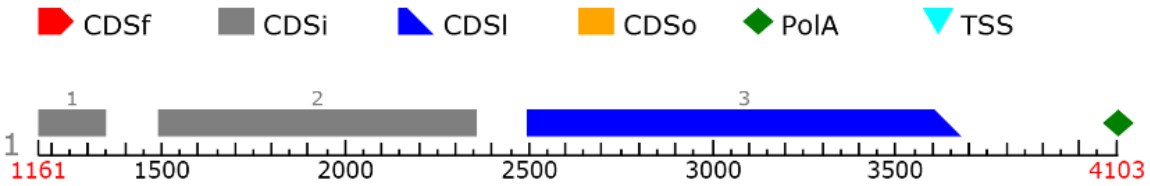
The initial *ab initio* annotation of Region 7 resulted in the prediction of several distinct gene models. Given the complexity of the hexaploid wheat genome, these models cannot be accepted as definitive without empirical support. Our validation strategy was designed to test each predicted mRNA and protein sequence individually against high-quality biological databases. A unique feature of our workflow was the parallel testing of "raw" predictions against "masked" versions. This allowed us to distinguish authentic genes from transposable element (TE) insertions that might have been erroneously modeled as genes by the prediction algorithms.

FGENESH comparison

As explained in the beginning, we mainly used Augustus as it is a software that is more conservative and will minimize false positives. However, it is interesting to use other software analysis to compare and observe if the genes predicted by Augustus will also be seen in the FGENESH software. For this comparison, we will only use the masked sequence analysis as we already know that the masked sequence will eliminate the possibility of transposable elements being identified as genes by the software.

Gene 1:

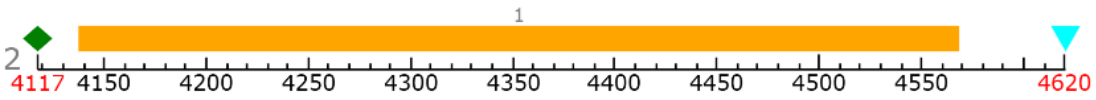
FGENESH 2.6 Prediction of potential genes in Triticum genomic DNA
Seq name: region7
Length of sequence: 15001
Number of predicted genes 4: in +chain 3, in -chain 1.
Number of predicted exons 12: in +chain 11, in -chain 1.
Positions of predicted genes and exons: Variant 1 from 1, Score:853.606641



1	+	1	CDSi	1161	-	1346	13.98	1161 - 1346 186
1	+	2	CDSi	1487	-	2356	283.42	1487 - 2356 870
1	+	3	CDSl	2492	-	3676	366.34	2492 - 3676 1185
1	+		PolA	4103			-3.01	

We observe that FGENESH predicts a total of 4 genes, 3 of them in the forward strand and 1 of them in the reverse strand. As expected, FGENESH predicts more genes than our Augustus tool. We observe that this first gene is composed of 3 exons, and it covers between position 1161 and position 4103. This is similar to the predicted gene 1 from Augustus, which covered from position 1305 until position 4122. There are some similarities, for example they both predict that the second exon will start at position 1487 and that there is a stop codon at position 3676, the difference is that FGENESH predicts 2 different exons whereas Augustus combines both into a single exon. Another similarity is that they both predict the stop codon at position 1346, they just differ on where the start codon is.

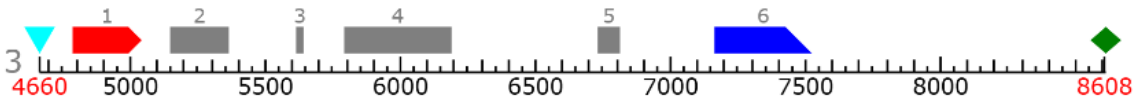
Gene 2:



2	-		PolA	4117		-6.61			
2	-	1	CDSo	4137	- 4568	61.58	4137	- 4568	432
2	-		TSS	4620		-13.31			

The second gene predicted by FGENESH is in the reverse strand and is also only one exon, the gene covers from position 4117 until position 4620. Augustus also predicts a gene in the reverse strand that is composed of a single exon, however, in a very different location (6774-7765).

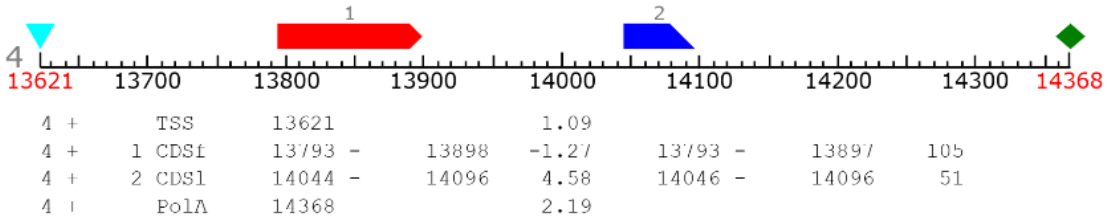
Gene 3:



3	+		TSS	4660		-11.21			
3	+	1	CDSf	4782	- 5038	32.46	4782	- 5036	255
3	+	2	CDSi	5143	- 5361	30.83	5144	- 5359	216
3	+	3	CDSi	5610	- 5637	-9.59	5611	- 5637	27
3	+	4	CDSi	5788	- 6185	59.90	5788	- 6183	396
3	+	5	CDSi	6727	- 6809	-7.70	6728	- 6808	81
3	+	6	CDSl	7159	- 7520	75.00	7161	- 7520	360
3	+		PolA	8608		2.19			

This gene is composed of 6 exons, is in the forward strand and the gene goes from position 4660 until position 8608. This gene is not predicted by Augustus entirely.

Gene 4:



4	+		TSS	13621		1.09		
4	+	1	CDSf	13793 -	13898	-1.27	13793 - 13897	105
4	+	2	CDSl	14044 -	14096	4.58	14046 - 14096	51
4	+		PolA	14368		2.19		

This gene is found almost at the end of our region, composed of 2 exons and goes from position 13621 until position 14368. This gene is also not predicted by Augustus in its entirety.

Overall, we see a variety of differences between both software predictions. In the first gene predicted by both we see various similarities, such as the intron-exon boundaries and the stop codon. However, there are some differences such as that FGENESH predicts three exons, the start of the first exon, etc. As expected, the FGENESH software is more aggressive in their predictions, partly due to it being normally used to predict full eukaryotic genomes and we are trying to predict a small region, which is why we see more genes and exons being predicted.

Multi-Locus Transcriptional Validation (BLASTn vs. TSA)

To validate the predicted gene models in Region 7, each mRNA sequence was independently queried against the Transcriptome Shotgun Assembly (TSA) database using BLASTn. This analysis was performed on both unmasked and masked versions of the predicted transcripts to evaluate whether transcriptional support persisted after the removal of repetitive sequences.

The Transcriptome Shotgun Assembly (TSA) database is a repository of publicly available, high-throughput transcript sequences from a wide variety of organisms. TSA contains assembled mRNA sequences derived from RNA sequencing experiments, providing comprehensive coverage of expressed genes across tissues, developmental stages, and experimental conditions. By querying predicted gene models against TSA using BLASTn, it is possible to validate transcriptional support, confirm exon-intron structures, and distinguish true protein-coding loci from repetitive or low-complexity sequences. This resource is particularly valuable for large and complex genomes, such as wheat, where direct experimental validation of all predicted transcripts is often impractical.

Strong transcriptional evidence was detected for Gene 1 and Gene 2, with BLASTn matches reaching up to 99.95% sequence identity and E-values of 0.0. These high-scoring alignments were observed consistently for both the unmasked and masked transcript versions, indicating that the transcriptional signal originates from unique genic sequences rather than from repetitive or transposable elements.

Gene 3 on the other hand, showed no detectable BLASTn alignment in either the unmasked or masked datasets. This result mirrors the Augustus predictions, where Gene 3 was not retained following repeat masking. The absence of transcriptomic support in both BLASTn analyses therefore reinforces the interpretation that this predicted locus does not correspond to a functional, transcribed gene.

For Gene 4, a BLASTn hit was detected only when the unmasked transcript was used, whereas no alignment was found after masking. This result matches the Augustus predictions, where Gene 4 was present in the unmasked sequence but disappeared once repetitive elements were removed. Taken together, these results suggest that the apparent transcriptional signal for Gene 4 originates from repetitive or transposable element sequences rather than from a true functional gene.

Importantly, the concordance between the ab initio predictions and transcript-based validation provides strong support for the final gene content of Region 7. Both Augustus and BLASTn independently converge on the same conclusion: only two gene models (Gene 1 and Gene 2) remain supported after accounting for repetitive DNA, while Genes 3 and 4 are likely artifacts of repeat-rich sequence context.

No	Description	Scientific_Name	Taxid	Max_Score	Total_Score	Query_Cover	E_value	Identity	Length	Accession
1	GG_102704_c0_g1_i1	Triticum aestivum	0	4072	4072	100%	0	99.95	2715	GKWO01368614.1
2	gene.191839.0.0	Triticum aestivum	0	4072	4072	100%	0	99.95	2936	GILY01020149.1
3	GG_102704_c0_g1_i2	Triticum aestivum	0	4043	4043	99%	0	99.95	2541	GKWO01368615.1
4	GG_35979_c0_g1_i1	Triticum aestivum	0	4019	4019	99%	0	99.95	2425	GKWO01132382.1
5	comp75839_c0_seq1 transcribed RNA sequence	Triticum aestivum	0	3993	3993	99%	0	99.54	2436	GEUX01058237.1
6	GG_106075_c0_g1_i4	Triticum aestivum	0	3312	3400	99%	0	94.03	2507	GKWO01379967.1
7	scaffold12062_size2563 transcribed RNA sequence	Triticum aestivum	0	3295	3383	99%	0	93.90	2563	GFFI01012061.1
8	GG_106075_c0_g1_i1	Triticum aestivum	0	3284	3372	99%	0	93.80	2507	GKWO01379966.1
9	gene.208233.0.0	Triticum aestivum	0	3284	3372	99%	0	93.80	2697	GILY01023572.1
10	evgBINPACKER.31463.7_Sumai:length_2581	Triticum aestivum	0	2796	2796	99%	0	89.94	2580	GIJS01012991.1

Table 1. BLASTn of Gene 1 (unmasked) against TSA confirms transcriptional support for the predicted gene model identified by Augustus.

No	Description	Scientific_Name	Taxid	Max_Score	Total_Score	Query_Cover	E_value	Identity	Length	Accession
1	GG_102704_c0_g1_i1	Triticum aestivum	0	4072	4072	100%	0	99.95	2715	GKWO01368614.1
2	gene.191839.0.0	Triticum aestivum	0	4072	4072	100%	0	99.95	2936	GILY01020149.1
3	GG_102704_c0_g1_i2	Triticum aestivum	0	4043	4043	99%	0	99.95	2541	GKWO01368615.1
4	GG_35979_c0_g1_i1	Triticum aestivum	0	4019	4019	99%	0	99.95	2425	GKWO01132382.1
5	comp75839_c0_seq1 transcribed RNA sequence	Triticum aestivum	0	3993	3993	99%	0	99.54	2436	GEUX01058237.1
6	GG_106075_c0_g1_i4	Triticum aestivum	0	3312	3400	99%	0	94.03	2507	GKWO01379967.1
7	scaffold12062_size2563 transcribed RNA sequence	Triticum aestivum	0	3295	3383	99%	0	93.90	2563	GFFI01012061.1
8	GG_106075_c0_g1_i1	Triticum aestivum	0	3284	3372	99%	0	93.80	2507	GKWO01379966.1
9	gene.208233.0.0	Triticum aestivum	0	3284	3372	99%	0	93.80	2697	GILY01023572.1
10	evgBINPACKER.31463.7_Sumai:length_2581	Triticum aestivum	0	2796	2796	99%	0	89.94	2580	GIJS01012991.1

Table 2. BLASTn of Gene 1 (masked) against TSA shows persistent transcriptional evidence after repeat masking, supporting the Augustus-predicted gene model.

No	Description	Scientific_Name	Taxid	Max_Score	Total_Score	Query_Cover	E_value	Identity	Length	Accession
1	comp72071_c0_seq1 transcribed RNA sequence	Triticum aestivum	0	760	760	71%	0e+00	99.05	990	GEUX01052893.1
2	GG_127776_c0_g1_i2	Triticum aestivum	0	747	1025	100%	0e+00	98.14	1651	GKWO01458962.1
3	GG_127776_c0_g1_i1	Triticum aestivum	0	747	1025	100%	0e+00	98.58	1571	GKWO01458961.1
4	TA_RNA_224655 transcribed RNA sequence	Triticum aestivum	0	747	747	72%	0e+00	98.14	785	GAJL01218132.1
5	cultivar Avocet R contig_122122	Triticum aestivum	0	532	604	51%	3e-148	98.36	793	GJAR01121435.1
6	GG_102715_c0_g1_i1	Triticum aestivum	0	448	448	41%	1e-122	100.00	353	GKWO01368652.1
7	NODE_128898_length_436	Triticum aestivum	0	322	322	30%	8e-85	98.89	436	HCED01128898.1
8	NODE_128898_length_436	Triticum aestivum	0	322	322	30%	8e-85	98.89	436	HCEC01128898.1
9	GG_102699_c0_g1_i1	Triticum aestivum	0	311	593	53%	2e-81	99.42	1145	GKWO01368602.1
10	NODE_188185_length_249	Triticum aestivum	0	278	278	29%	2e-71	95.93	249	HCED01188185.1

Table 3. BLASTn of Gene 2 (unmasked) against TSA indicates strong transcriptional support for the Augustus-predicted gene model.

No	Description	Scientific_Name	Taxid	Max_Score	Total_Score	Query_Cover	E_value	Identity	Length	Accession
1	comp72071_c0_seq1 transcribed RNA sequence	Triticum aestivum	0	760	760	71%	0e+00	99.05	990	GEUX01052893.1
2	GG_127776_c0_g1_i2	Triticum aestivum	0	747	1025	100%	0e+00	98.14	1651	GKWO01458962.1
3	GG_127776_c0_g1_i1	Triticum aestivum	0	747	1025	100%	0e+00	98.58	1571	GKWO01458961.1
4	TA_RNA_224655 transcribed RNA sequence	Triticum aestivum	0	747	747	72%	0e+00	98.14	785	GAJL01218132.1
5	cultivar Avocet R contig_122122	Triticum aestivum	0	532	604	51%	3e-148	98.36	793	GJAR01121435.1
6	GG_102715_c0_g1_i1	Triticum aestivum	0	448	448	41%	1e-122	100.00	353	GKWO01368652.1
7	NODE_128898_length_436	Triticum aestivum	0	322	322	30%	8e-85	98.89	436	HCED01128898.1
8	NODE_128898_length_436	Triticum aestivum	0	322	322	30%	8e-85	98.89	436	HCEC01128898.1
9	GG_102699_c0_g1_i1	Triticum aestivum	0	311	593	53%	2e-81	99.42	1145	GKWO01368602.1
10	NODE_188185_length_249	Triticum aestivum	0	278	278	29%	2e-71	95.93	249	HCED01188185.1

Table 4. BLASTn of Gene 2 (masked) against TSA shows consistent transcriptional evidence after repeat masking, confirming the Augustus prediction.

No	Description	Scientific_Name	Taxid	Max_Score	Total_Score	Query_Cover	E_value	Identity	Length	Accession
1	TSA: Triticum aestivum scaffold122295_size674 transcribed RNA sequence	Triticum aestivum	0	560	560	88%	1e-156	87.12	674	GFFI01122270.1
2	TSA: Triticum aestivum scaffold102255_size809 transcribed RNA sequence	Triticum aestivum	0	542	542	83%	5e-151	87.04	809	GFFI01102237.1
3	TSA: Triticum aestivum scaffold89973_size911 transcribed RNA sequence	Triticum aestivum	0	520	520	76%	3e-144	88.79	911	GFFI01089959.1
4	Triticum aestivum GG_33967_c0_g1_i1	Triticum aestivum	0	372	372	45%	7e-100	92.37	368	GKWO01124597.1
5	TSA: Triticum aestivum tid185308_c0_seq1 transcribed RNA sequence	Triticum aestivum	0	230	230	31%	5e-57	89.62	234	GEWU01338768.1
6	Triticum aestivum GG_52879_c21_g1_i1	Triticum aestivum	0	226	226	32%	6e-56	88.77	441	GKWO01192222.1
7	Triticum aestivum GG_122906_c0_g1_i2	Triticum aestivum	0	224	224	34%	2e-55	87.82	3365	GKWO01441379.1
8	Triticum aestivum gene.294229.0.0	Triticum aestivum	0	224	224	34%	2e-55	87.82	3697	GILY01042030.1
9	TSA: Triticum aestivum scaffold96462_size856 transcribed RNA sequence	Triticum aestivum	0	224	224	34%	2e-55	87.82	856	GFFI01096447.1
10	Triticum aestivum evgTaSumailDBA_57_135662:length_1078	Triticum aestivum	0	206	206	34%	8e-50	85.86	1077	GIJS01123902.1

Table 5. BLASTn of Gene 4 (masked) against TSA reveals absence of transcriptional support after repeat masking, indicating that the Augustus-predicted gene model may correspond to repetitive elements.

Comparative Proteomic Validation (BLASTp vs. NR & SwissProt)

Following transcriptional validation, we evaluated whether the predicted gene models correspond to biologically meaningful proteins by performing BLASTp searches against the clustered NR and SwissProt databases. Each predicted protein was analyzed in both unmasked and masked contexts, when available, to determine whether the encoded proteins represent genuine coding loci or are instead derived from repetitive or transposable-element–associated sequences.

Protein 1 showed robust and consistent BLASTp support in both its unmasked and masked forms (Table 6; Table 7). In both cases, the top hit corresponded to an E3 ubiquitin-protein ligase WAV3, with a query coverage of 100%, an E-value of 0.0, and 97.28% sequence identity to homologs from *Aegilops tauschii*. Additional high-scoring matches were observed across multiple Triticeae species, including *Triticum turgidum*, *Triticum dicoccoides*, and *Hordeum vulgare*, consistently spanning the full protein length of approximately 735 amino acids. The persistence of these high-confidence hits after repeat masking demonstrates that Protein 1 is not driven by repetitive DNA and represents a conserved wheat protein-coding gene. The broad taxonomic distribution and high conservation strongly support its functional annotation as an E3 ubiquitin ligase, a protein family known to play key roles in plant regulatory and developmental pathways.

Protein 3, a short predicted protein of 88 amino acids, yielded detectable but weaker BLASTp hits exclusively in the unmasked sequence (Table 8). Significant alignments were obtained against hypothetical proteins from *Triticum aestivum* and *Hordeum vulgare*, with E-values ranging from 8e−10 to 3e−4, sequence identities between 45% and 59%, and query coverage around 60–65%. The absence of hits for the masked version, combined with the short length of the protein and its limited conservation, suggests that Protein 3 may correspond to a low-confidence or lineage-specific coding sequence, or potentially a partial ORF overlapping repetitive regions.

While some protein-level similarity exists, the overall evidence remains weak compared to Protein 1 and should therefore be interpreted with caution.

Protein 4 produced significant BLASTp hits only in the unmasked sequence (Table 9), with no detectable homologs after masking. The unmasked protein aligned to hypothetical wheat proteins with E-values as low as 4e−15, but with moderate query coverage ranging from 21% to 47% and sequence identity between 59% and 83%. These alignments were restricted to relatively short regions of the protein. This behavior mirrors what was observed at both the Augustus prediction and BLASTn transcript validation levels, where Gene 4 was present only in the unmasked sequence and disappeared after repeat masking. The aligned regions were enriched in glycine- and arginine-rich low-complexity motifs, a common feature of transposable element–associated sequences in wheat. Together, these observations indicate that Protein 4 does not represent a true protein-coding gene, but rather reflects a false-positive prediction driven by repetitive genomic elements.

No significant BLASTp hits were obtained for the predicted protein corresponding to Gene 2 in either masked or unmasked datasets. The absence of protein-level support is consistent with the lack of transcriptomic evidence and further suggests that this predicted gene model does not encode a stable or conserved protein product.

Overall, protein-level validation supports only one high-confidence gene in Region 7, corresponding to Gene 1, which remains stable in both masked and unmasked analyses. Proteins corresponding to Genes 3 and 4 show limited or masking-sensitive support, while Gene 2 lacks detectable protein homology altogether. This pattern is fully consistent with the Augustus predictions, in which repeat masking reduced the number of predicted genes from four to two. The convergence of gene prediction, transcript validation, and protein homology analyses demonstrates that repeat masking is essential for eliminating false positives in the wheat genome and for producing a reliable final annotation.

No	Cluster_Representative	Members	Taxa	Scientific_Name	Taxid	Max_Score	Total_Score	Query_Cover	E_value	Identity	Length	Accession
1	E3 ubiquitin-protein ligase WAV3 isoform X3 [Aegilops tauschii]	13	5	Triticinae	1648030	1436	1436	100%	0	97.28	723	XP_073353260.1
2	unnamed protein product [Triticum turgidum subsp. durum]	1	1	Triticum turgidum subsp. durum	4567	1219	1219	99%	0	90.22	731	VAI25072.1
3	hypothetical protein ZWY2020_035386 [Hordeum vulgare]	3	2	Hordeum vulgare	4513	1113	1113	100%	0	86.77	714	KAI4995483.1
4	unnamed protein product [Triticum turgidum subsp. durum]	1	1	Triticum turgidum subsp. durum	4567	1062	1062	81%	0	92.82	604	VAI25073.1
5	E3 ubiquitin-protein ligase WAV3-like [Triticum dicoccoides]	2	2	Triticum dicoccoides	4564	962	962	99%	0	75.38	692	XP_037433296.1
6	unnamed protein product [Triticum turgidum subsp. durum]	1	1	Triticum turgidum subsp. durum	4567	953	953	79%	0	90.65	775	VAI10714.1
7	hypothetical protein ZWY2020_035380 [Hordeum vulgare]	1	1	Hordeum vulgare	4513	908	1023	87%	0	88.74	671	KAI4995477.1
8	E3 ubiquitin-protein ligase WAV3-like [Lolium rigidum]	5	4	Lolinae	640630	875	875	99%	0	70.05	659	XP_047083527.1
9	unnamed protein product [Alopecurus aequalis]	1	1	Alopecurus aequalis	114194	859	859	98%	0	69.55	637	CAM0870680.1
10	hypothetical protein ACUV84_007391 [Puccinellia chinampoensis]	1	1	Puccinellia chinampoensis	428215	836	836	99%	0	68.84	654	KAM3064478.1

Table 6. BLASTp of Protein 1 (unmasked) against NR shows full-length, high-identity matches to E3 ubiquitin-protein ligase WAV3 homologs (E = 0.0).

No	Cluster_Representative	Members	Taxa	Scientific_Name	Ancestor	Taxid	Max_Score	Total_Score	Query_Cover	E_value	Identity	Length	Accession
1	E3 ubiquitin-protein ligase WAV3 isoform X3 [Aegilops tauschii]	13	5	Triticinae	monocots	1648030	1436	1436	100%	0	97.28	723	XP_073353260.1
2	unnamed protein product [Triticum turgidum subsp. durum]	1	1	Triticum turgidum	durum wheat	4567	1219	1219	99%	0	90.22	731	VAI25072.1
3	hypothetical protein ZWY2020_035386 [Hordeum vulgare]	3	2	Hordeum vulgare	barley	4513	1113	1113	100%	0	86.77	714	KAI4995483.1
4	unnamed protein product [Triticum turgidum subsp. durum]	1	1	Triticum turgidum	durum wheat	4567	1062	1062	81%	0	92.82	604	VAI25073.1
5	E3 ubiquitin-protein ligase WAV3-like [Triticum dicoccoides]	2	2	Triticum	monocots	4564	962	962	99%	0	75.38	692	XP_037433296.1
6	unnamed protein product [Triticum turgidum subsp. durum]	1	1	Triticum turgidum	durum wheat	4567	953	953	79%	0	90.65	775	VAI10714.1
7	hypothetical protein ZWY2020_035380 [Hordeum vulgare]	1	1	Hordeum vulgare	barley	4513	908	1023	87%	0	88.74	671	KAI4995477.1
8	E3 ubiquitin-protein ligase WAV3-like [Lolium rigidum]	5	4	Loliinae	monocots	640630	875	875	99%	0	70.05	659	XP_047083527.1
9	unnamed protein product [Alopecurus aequalis]	1	1	Alopecurus aequalis	monocots	114194	859	859	98%	0	69.55	637	CAM0870680.1
10	hypothetical protein ACUV84_007391 [Puccinellia chinampoensis]	1	1	Puccinellia chinampoensis	monocots	428215	836	836	99%	0	68.84	654	KAM3064478.1

Table 7. BLASTp of Protein 1 (masked) against NR yields the same high-scoring hits as the unmasked sequence, confirming a non-repetitive protein-coding gene.

No	Cluster_Representative	Mem	Taxa	Scientific_Name	Ancestor	Taxid	Score	Cover	E_value	Ident	Len	Accession
1	hypothetical protein VPH35_124735 [Triticum aestivum]	1	1	T. aestivum	bread wheat	4565	62.8	65%	8e-10	52.6	158	XBI40084.1
2	hypothetical protein VPH35_070349 [Triticum aestivum]	1	1	T. aestivum	bread wheat	4565	50.1	61%	3e-05	59.3	122	XBI77192.1
3	hypothetical protein D1007_26523 [Hordeum vulgare]	1	1	H. vulgare	barley	4513	48.1	60%	3e-04	45.3	193	KAE8798229.1

Table 8. BLASTp of Protein 3 (unmasked) against NR returns only short, partial matches to hypothetical proteins.

No	Cluster_Representative	Mem	Taxa	Scientific_Name	Ancestor	Taxid	Score	Cover	E_value	Ident	Len	Accession
1	hypothetical protein CFC21_096101 [Triticum aestivum]	4	2	Triticum	monocots	4564	79.3	47%	4e-15	59.41	133	KAF7093709.1
2	hypothetical protein VPH35_103550 [Triticum aestivum]	1	1	T. aestivum	bread wheat	4565	70.5	21%	8e-12	82.93	132	XBH77003.1

Table 9. BLASTp of Protein 4 (unmasked) against NR identifies partial matches to hypothetical wheat proteins, with no significant hits after masking.

Region-Wide Sensitivity Analysis (BLASTx)

To further ensure the completeness of the predicted gene models in Region 7, a BLASTX search was performed across the entire ~15 kb genomic sequence. This approach translates the DNA in all six reading frames and compares the resulting protein sequences against the SwissProt database, allowing independent detection of coding potential beyond the ab initio predictions.

The BLASTX analysis revealed several high-confidence alignments (E-values ≤ 4e−27) corresponding precisely to the coordinates of previously predicted genes. Notably, strong hits were detected for Protein 1, consistent with its transcriptomic and BLASTp validation. These alignments included E3 ubiquitin-protein ligase homologs from **Arabidopsis thaliana**, **Oryza sativa**, and related monocot species, confirming the coding potential and conserved function of this locus (see Table 10).

For the remaining predicted genes, BLASTX results mirrored the patterns observed in BLASTn and BLASTp analyses. Genes 3 and 4 yielded weak or partial alignments,

restricted to repetitive or low-complexity regions, while Gene 2 showed no detectable protein homology. Importantly, no additional coding regions were identified outside the predicted loci, indicating that the ab initio predictions captured all biologically meaningful genes within this genomic segment.

Overall, the BLASTX results provide **an independent, genome-wide confirmation** that the final annotation of Region 7 is both accurate and complete. The absence of significant hits in intergenic regions supports the conclusion that the predicted coding sequences are genuine and not artifacts of repetitive DNA or spurious open reading frames.

No	Description	Scientific_Name	Common_Name	Taxid	Max_Score	Query_Cover	E_value	Identity	Accession
1	E3 ubiquitin-protein ligase WAV3 [Arabidopsis thaliana]	Arabidopsis thaliana	thale cress	3702	125.0	7%	4e-27	36.71	Q9LTA6.1
2	E3 ubiquitin-protein ligase WAVH1 [Arabidopsis thaliana]	Arabidopsis thaliana	thale cress	3702	109.0	7%	3e-22	29.34	Q9ZQ46.1
3	Probable E3 ubiquitin-protein ligase EDA40 [Arabidopsis thaliana]	Arabidopsis thaliana	thale cress	3702	96.3	7%	4e-18	31.72	F4JSV3.1
4	Uncharacterized protein sll0103 [Synechocystis sp.]	Synechocystis sp.	NA	1111708	91.3	5%	3e-17	30.88	Q55874.1
5	Inter-alpha-trypsin inhibitor heavy chain H4 [Homo sapiens]	Homo sapiens	human	9606	72.8	4%	8e-11	29.72	Q14624.4
6	Calcium-activated chloride channel regulator 1 [Bos taurus]	Bos taurus	cattle	9913	49.7	2%	8e-04	31.78	P54281.1
7	Calcium-activated chloride channel regulator 1 [Equus caballus]	Equus caballus	horse	9796	48.9	4%	0.001	27.46	Q2TU62.1
8	Probable E3 ubiquitin-protein ligase WAVH2 [Arabidopsis thaliana]	Arabidopsis thaliana	thale cress	3702	47.8	1%	0.003	40.00	Q0WQX9.1
9	E3 ubiquitin-protein ligase IPI1 [Oryza sativa]	Oryza sativa	rice	39947	46.6	1%	0.006	48.94	Q5Z8R1.1
10	RING-H2 finger protein ATL18 [Arabidopsis thaliana]	Arabidopsis thaliana	thale cress	3702	43.1	1%	0.010	43.48	Q9SZL4.1

Table 10. BLASTX of Region 7 against SwissProt shows high-confidence alignments for Protein 1 and limited or partial matches for other predicted gene models, confirming coding potential and correspondence with previously annotated genes.

Transposable Elements Validation

Transposable element (TE) annotation and validation are critical in wheat due to the extreme abundance of repetitive sequences and their tendency to generate false gene predictions. To characterize the TE content of Region 7 and evaluate its impact on gene annotation, we combined homology-based searches against the TREP database, URGI database, dotplot structural analysis, and genome visualization in Artemis using Censor, Augustus, and FGENESH annotation tracks.

BLASTn searches against the TREP database revealed multiple high-confidence TE matches distributed across the entire 15 kb region, corresponding to both Class I retrotransposons (e.g. LINE/L1 and Copia elements) and Class II DNA transposons

(e.g. MuDR/Mutator, EnSpm/CACTA, and Mariner families) (Table 11). The alignments displayed very low E-values and high sequence identity, confirming that these regions represent authentic TE-derived sequences rather than spurious low-complexity matches. Based on the alignment against the URGI database (Table 12), several Censor-annotated transposable elements were also identified by BLAST. The Mariner-N10_TAe element was detected by BLAST as DTX-incomp_3b_Itr1_300M-B-G24679-Map10_revers_917, with 89.247% sequence identity, while Censor located this element at sequence positions 6918–7095 and 6424–6520. Similarly, the MuDR-59N1_TAe element corresponded to the BLAST hit RIX-comp_3b_Itr1_300M-B-R4417-Map4_1468, showing 87.546% identity, and was identified by Censor at positions 9798–10558. For Copia-82_TAe, both the LTR and internal regions were detected. BLAST returned several hits corresponding to RLX-comp_3b_Itr1_300M-L-B5896-Map1_reversed_3331, with identities reaching up to 94.372% for the region starting at position 11389. Censor annotated the LTR region at positions 11389–11650 and the internal regions at positions 11844–13112 and 14634–14995. In addition, both tools successfully identified the same EnSpm insertions, despite differences in the specific database identifiers used. Overall, BLAST supports the CENSOR predictions.

Censor annotation indicates that approximately **31% of Region 7 is composed of transposable element fragments.** Although this value is lower than the ~80–85% TE content reported at the whole-genome scale for wheat, this is biologically coherent, as TE density is highly heterogeneous, and gene-rich regions typically show reduced repeat content.

Dotplot comparisons between Region 7 and representative URGI elements further supported the BLAST results by revealing extended diagonal similarity blocks, often interrupted by gaps, which is characteristic of fragmented or degenerated TE insertions commonly observed in large plant genomes. In particular, comparison with the LINE element L1-384 showed a discontinuous diagonal typical of ancient non-LTR retrotransposon fragments (Figure 1), while comparison with the MuDR-N38 DNA transposon revealed conserved blocks corresponding to Mutator-like insertions dispersed across the region (Figure 2). These structural patterns confirm that the detected similarities correspond to genuine transposable element remnants embedded in Region 7.

Artemis visualization provided direct spatial evidence for the interaction between TEs and predicted genes. When displaying the Censor TE track alone, numerous TE fragments spanning several families (MuDR, EnSpm, Mariner, L1, Copia) were observed throughout Region 7 (Figure 3). When the unmasked Augustus predictions were overlaid with the TE track, several CDS features as well as start and stop codon annotations were found to fall entirely within TE coordinates (Figure 4). This

indicates that part of the Augustus gene models, particularly those that disappear after repeat masking, are driven by TE-derived sequence rather than by true host genes. In contrast, when the unmasked FGENESH predictions were visualized together with the TE track, little to no systematic overlap between CDS features and TE intervals was observed (Figure 5), suggesting that FGENESH is more conservative in repetitive regions and less prone to interpreting TE fragments as protein-coding loci.

Together, BLASTn homology, dotplot structure, and Artemis co-localization demonstrate that a substantial fraction of Region 7 consists of transposable element sequences and that these elements directly account for the spurious gene predictions observed in unmasked analyses. This TE validation therefore explains the reduction in predicted gene number after masking and highlights the necessity of integrating repeat annotation with gene prediction in the wheat genome.

KEY	Hit_count	Total_aligned_bp	Mean_identity_pct	Max_identity_pct	Best_Evalue	Region_min	Region_max	urgi_subject	TE_class	TE_family	total_overlap_bp
KEY=1030	5	1528	86.42	89.189	0.0e+00	11388	14966	_300M-L-B5896-Ma	etrotransposon (LTF	LTR (RLX)	1735
KEY=436	6	1473	88.62	90.141	0.0e+00	11388	14966	_300M-L-B5896-Ma	etrotransposon (LTF	LTR (RLX)	1392
KEY=3916	2	393	88.4	88.889	2.2e-108	9798	10133	_Itr1_300M-B-R441	otransposon (non-L	RIX	324
KEY=1387	2	218	85.17	89.831	3.3e-22	8708	8851	_3b_Itr1_300M-B-F	DNA transposon	DTX	202
KEY=1646	1	72	84.72	84.722	2.0e-09	6651	6714	_Itr1_300M-B-G10(	DNA transposon	DTX	64

Table 11. BLASTn summary of transposable element matches in Region 7 against TREP, grouped by TE class/family label with alignment statistics.

region7	RLX-comp_3b_Itr1_300M-L-B5896-Map1_reversed_3331	84.229	1414	141	53	12762	14161	1237	2582	0.0	1301
region7	RLX-comp_3b_Itr1_300M-L-B5896-Map1_reversed_3331	93.935	643	34	4	14363	15001	3171	3812	0.0	966
region7	RLX-comp_3b_Itr1_300M-L-B5896-Map1_reversed_3331	94.372	231	10	3	11389	11618	6	234	8.55e-95	351
region7	RLX-comp_3b_Itr1_300M-L-B5896-Map1_reversed_3331	92.241	232	15	3	11388	11618	5105	5334	5.18e-87	326
region7	RLX-comp_3b_Itr1_300M-L-B5896-Map1_reversed_3331	87.912	91	9	2	11920	12009	483	572	7.35e-21	106
region7	RLX-comp_3b_Itr1_300M-L-B5853-Map1_3315 84.449	1061	114	35	13119	14161	1704	2731	0.0	998	
region7	RLX-comp_3b_Itr1_300M-L-B5853-Map1_3315 94.264	645	25	3	14360	15001	3316	3951	0.0	976	
region7	RLX-comp_3b_Itr1_300M-L-B5853-Map1_3315 91.358	162	11	2	14208	14369	2744	2902	9.04e-55	219	
region7	RLX-comp_3b_Itr1_300M-L-B5791-Map1_reversed_3298	87.970	798	66	19	13511	14307	2164	2932	0.0	915
region7	RLX-comp_3b_Itr1_300M-L-B5791-Map1_reversed_3298	88.571	350	29	11	13075	13417	1773	2118	1.07e-113	414
region7	RIX-comp_3b_Itr1_300M-B-R4417-Map4_1468 87.546	273	32	2	9800	10070	2073	1801	1.12e-83	315	
region7	DTX-incomp_3b_Itr1_300M-B-G24679-Map10_revers_917	89.247	186	18	2	6918	7103	10622	10439	1.16e-58	231
region7	RLX-incomp_3b_Itr1_300M-L-B5100-Map1_6029	88.827	179	19	1	6918	7096	1081	904	9.04e-55	219
region7	RLX-incomp_3b_Itr1_300M-B-R3425-Map5_reversed_5417	93.902	82	5	0	13502	13583	82	1	2.03e-26	124
region7	DTX-incomp-chim_3b_Itr1_300M-L-B4539-Map1_rev_565	87.156	109	14	0	697	805	3049	2941	2.03e-26	124
region7	RLX-comp_3b_Itr1_300M-L-B5253-Map1_3177 93.590	78	3	2	11932	12009	479	554	1.22e-23	115	
region7	PotentialHostGene_3b_Itr1_300M-B-R1379-Map3_1349	79.630	162	19	4	8706	8853	15078	14917	2.64e-20	104
region7	DTX-incomp_3b_Itr1_300M-B-G10654-Map10_715	92.958	71	4	1	6651	6720	580	650	9.50e-20	102
region7	DTX-incomp_3b_Itr1_300M-B-R5904-Map4_1149	91.429	70	6	0	6651	6720	581	650	4.42e-18	97.1
region7	RLX-incomp_3b_Itr1_300M-L-B4071-Map1_reversed_5907	81.982	111	20	0	695	805	2431	2541	1.59e-17	95.3
region7	DTX-incomp-chim_3b_Itr1_300M-B-P475.0-Map3_141	77.778	162	22	3	8706	8853	2205	2366	2.66e-15	87.9
region7	RIX-incomp-chim_3b_Itr1_300M-B-G25269-Map3_1492 86.250	80	9	2	8775	8853	7321	7399	9.57e-15	86.1	
region7	RXX-LARD_3b_Itr1_300M-L-B6477-Map1_6507 84.146	82	12	1	10267	10348	2456	2376	1.60e-12	78.7	
region7	RIX-comp-chim_3b_Itr1_300M-B-R3999-Map6_rever_1394	83.544	79	13	0	8775	8853	6591	6513	2.07e-11	75.0
region7	RLX-incomp-chim_3b_Itr1_300M-B-G3155-Map8_3567	93.750	48	3	0	8706	8753	15591	15638	7.45e-11	73.1
region7	DTX-incomp_3b_Itr1_300M-B-G17249-Map3_reverse_767	93.750	48	3	0	8706	8753	833	880	7.45e-11	73.1
region7	DTX-incomp_3b_Itr1_300M-B-R5798-Map3_1145	84.615	78	4	5	6651	6720	574	651	2.68e-10	71.3
region7	DTX-incomp_3b_Itr1_300M-B-R1487-Map4_reversed_1022	76.220	164	21	11	8706	8853	751	590	2.68e-10	71.3
region7	RLX-incomp-chim_3b_Itr1_300M-B-G24806-Map4_re_3540	91.667	48	4	0	8706	8753	7277	7324	3.47e-09	67.6
region7	RLX-incomp-chim_3b_Itr1_300M-L-B1264-Map1_3703	85.714	63	7	2	8794	8854	3769	3831	1.25e-08	65.8
region7	DTX-incomp_3b_Itr1_300M-B-G22562-Map9_reverse_845	84.932	73	2	4	6651	6714	591	663	1.25e-08	65.8
region7	DTX-incomp-chim_3b_Itr1_300M-L-B2072-Map1_rev_302	85.507	69	2	2	6652	6720	2633	2693	1.25e-08	65.8
region7	DTX-incomp-chim_3b_Itr1_300M-L-B157-Map1_239	91.489	47	3	1	6651	6696	13183	13229	4.48e-08	63.9
region7	DTX-incomp-chim_3b_Itr1_300M-L-B1676-Map1_245	91.304	46	3	1	6651	6696	1366	1410	1.61e-07	62.1
region7	DTX-incomp_3b_Itr1_300M-B-G22554-Map20_revers_844	100.000	32	0	0	6651	6682	585	616	5.80e-07	60.2
region7	DTX-incomp_3b_Itr1_300M-B-G16806-Map7_reverse_761	100.000	32	0	0	6651	6682	512	543	5.80e-07	60.2

Table 12. BLASTn summary of transposable element matches in Region 7 against URGI, grouped by TE class/family label with alignment statistics.

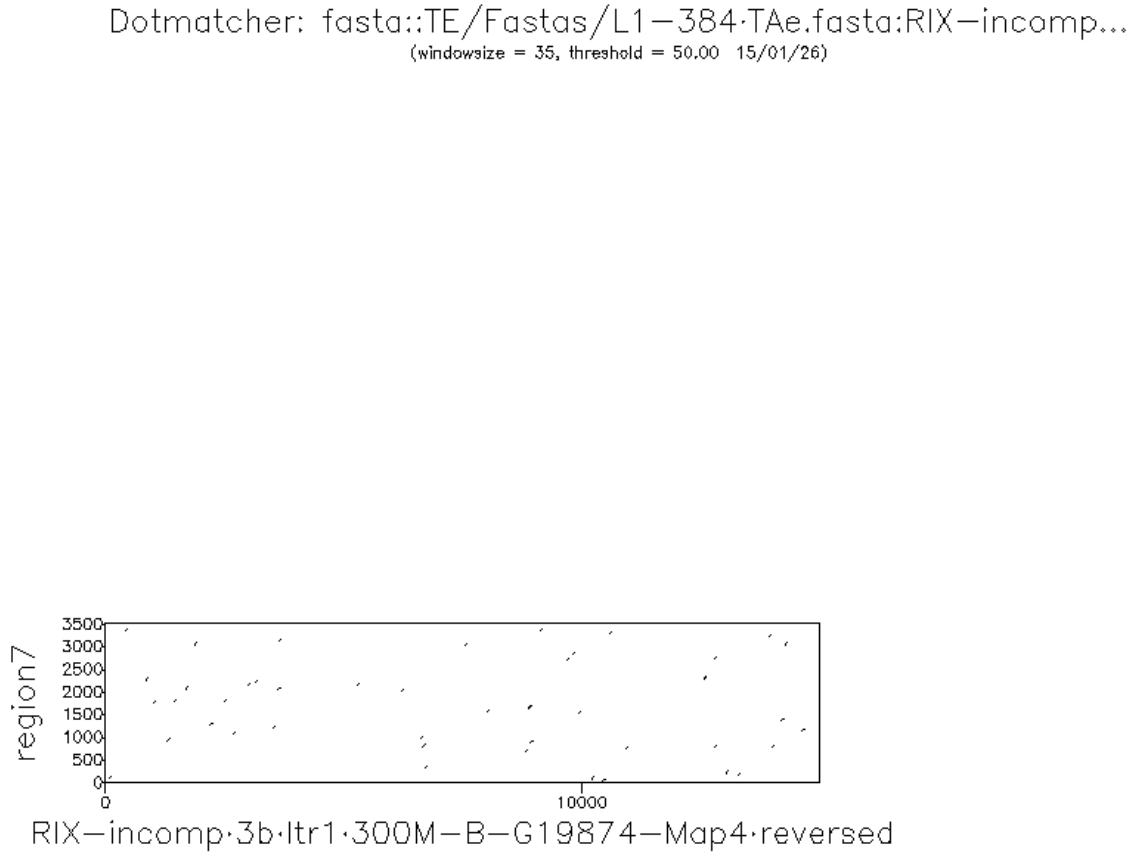


Figure 1. Dotplot showing sequence similarity between Region 7 and the LINE retrotransposon L1-384 (RIX family). The fragmented diagonal pattern indicates multiple degenerated non-LTR retrotransposon insertions within the region.

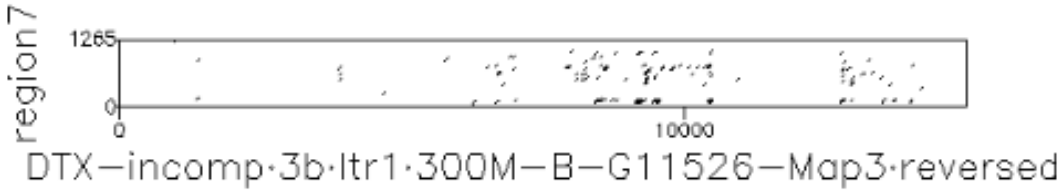


Figure 2. Dotplot showing sequence similarity between Region 7 and the MuDR-N38 DNA transposon (Mutator superfamily). Conserved blocks reflect fragmented DNA transposon insertions distributed across the region.

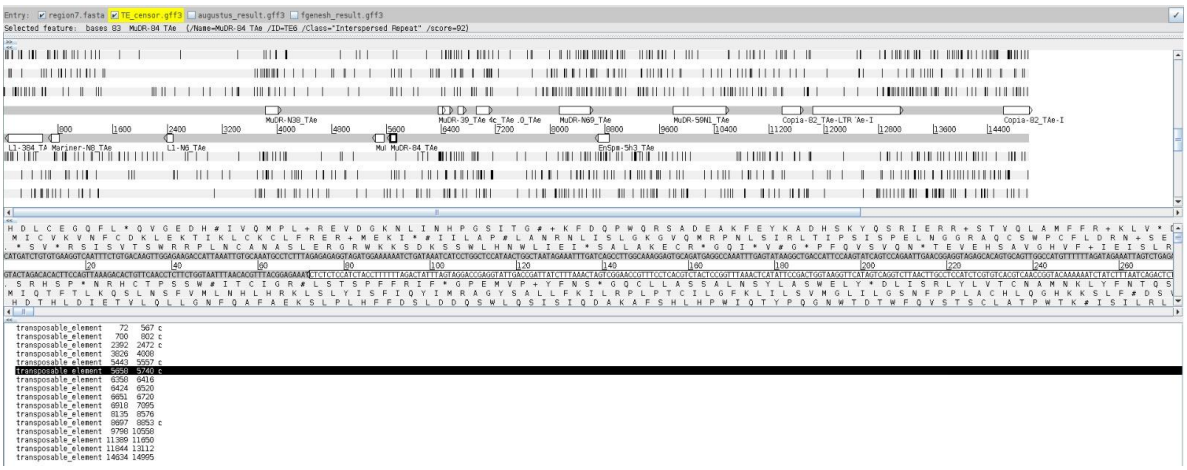


Figure 3. Artemis view of Region 7 displaying Censor TE annotations across the region.

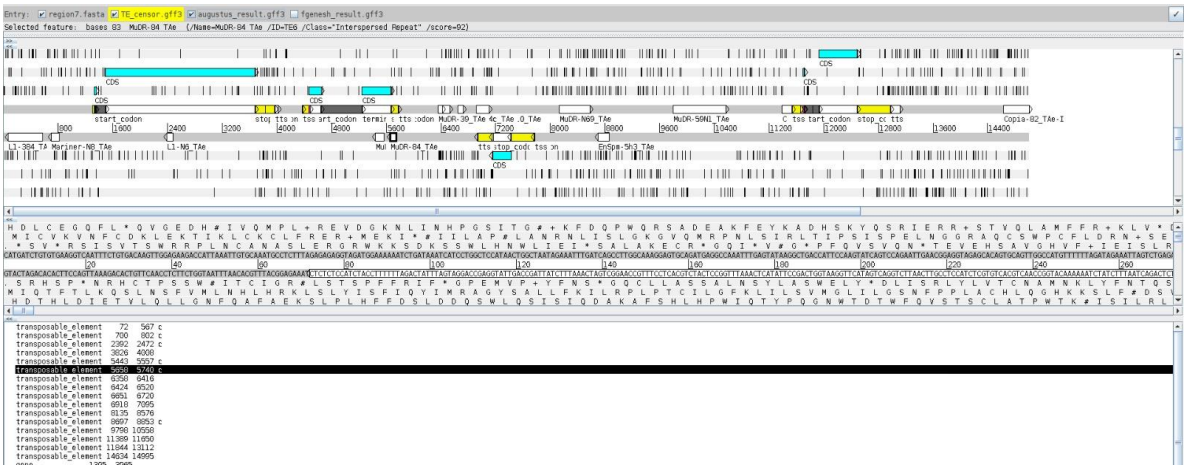


Figure 4. Artemis view of Region 7 with Censor TE annotations and unmasked Augustus predictions; multiple CDS/start/stop features overlap TE intervals.

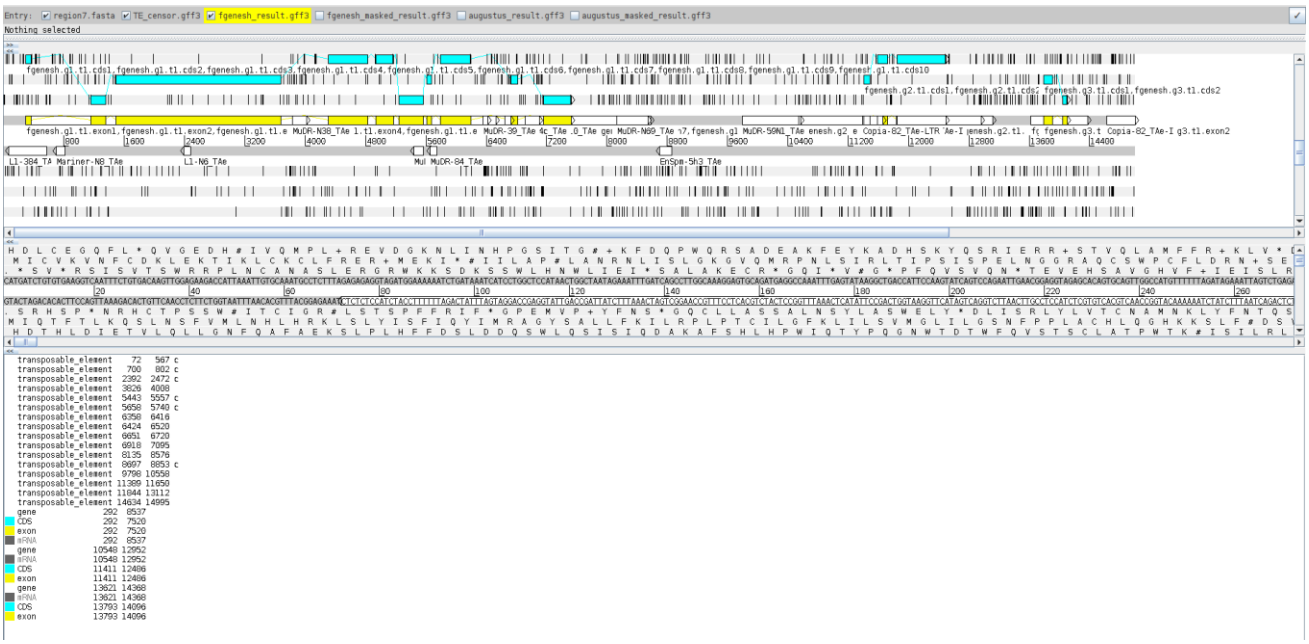


Figure 5. Artemis view of Region 7 with Censor TE annotations and unmasked FGENESH predictions; fewer coding features overlap TE coordinates than observed for Augustus.

ARTEMIS

Comparison of gene predictions in Artemis, a genome visualization and annotation platform that allows simultaneous inspection of sequence features, gene models, and evidence tracks, reveals clear differences between Augustus and FGENESH and highlights the impact of repeat masking. In the unmasked sequence, Augustus predicts four genes (Figure 6), and FGENESH predicts similar loci but with additional CDS fragments and alternative structures (Figure 9). After masking, Augustus becomes highly conservative and retains only two gene models (Figure 7), corresponding to the loci supported by transcript and protein evidence. In contrast, FGENESH still predicts extra CDS and partial gene structures in the masked sequence (Figure 8), indicating a higher sensitivity but also a greater tendency to overpredict in repetitive or low-complexity regions. Together, these results support the conclusion that two genes represent the most reliable annotations in Region 7, whereas the additional models observed in unmasked analyses and in FGENESH are likely influenced by repetitive sequences.

Figure 6. Artemis view of Region 7 showing unmasked Augustus predictions. Four gene models are detected, with multiple CDS, start codons, and stop codons distributed across the region.

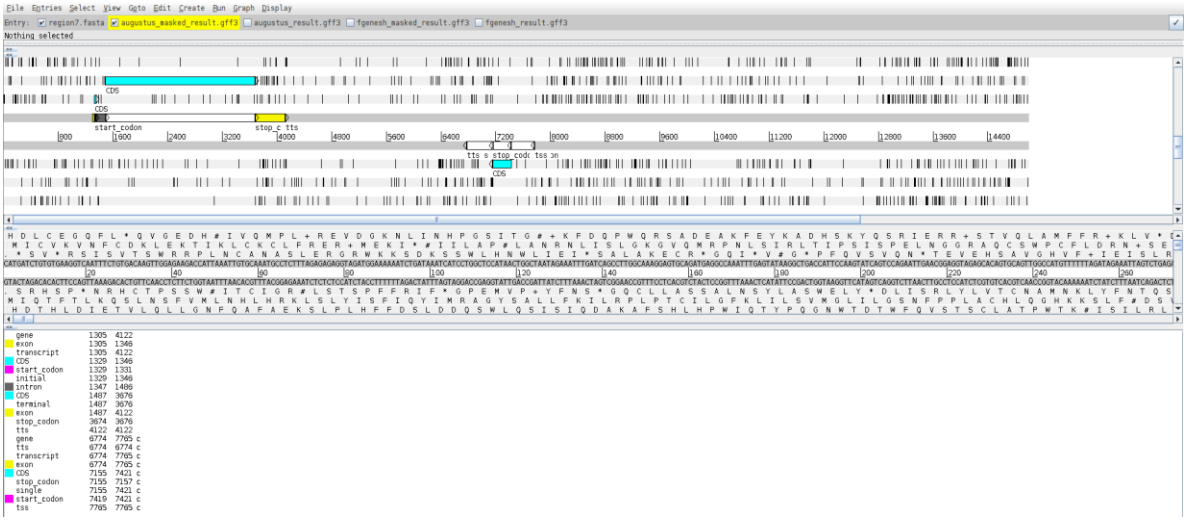


Figure 7. Artemis view of Region 7 showing masked Augustus predictions. Only two gene models remain after repeat masking, while the other predicted genes disappear, indicating that they were supported mainly by repetitive sequences.

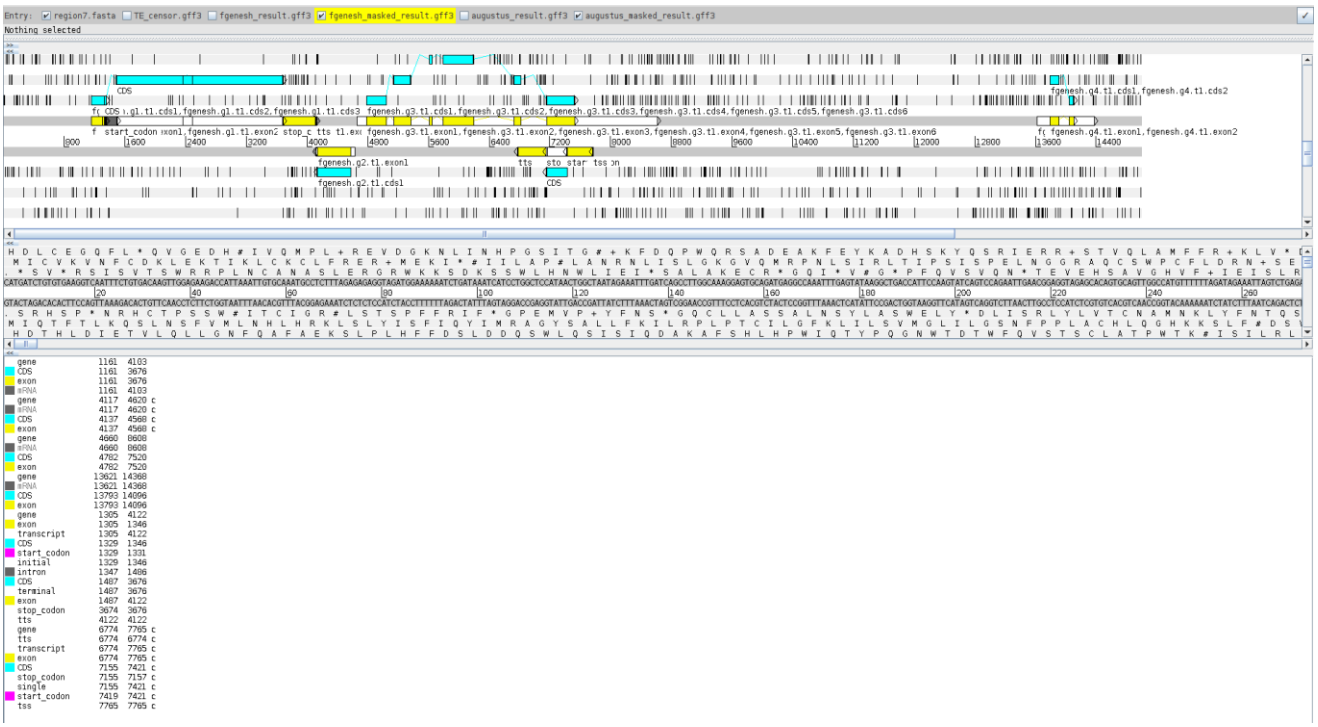


Figure 8. Artemis view of Region 7 showing masked Augustus and masked FGENESH predictions displayed simultaneously. Augustus retains two genes, whereas FGENESH still predicts additional CDS fragments and partial gene structures, reflecting a higher sensitivity of FGENESH in repeat-rich regions.

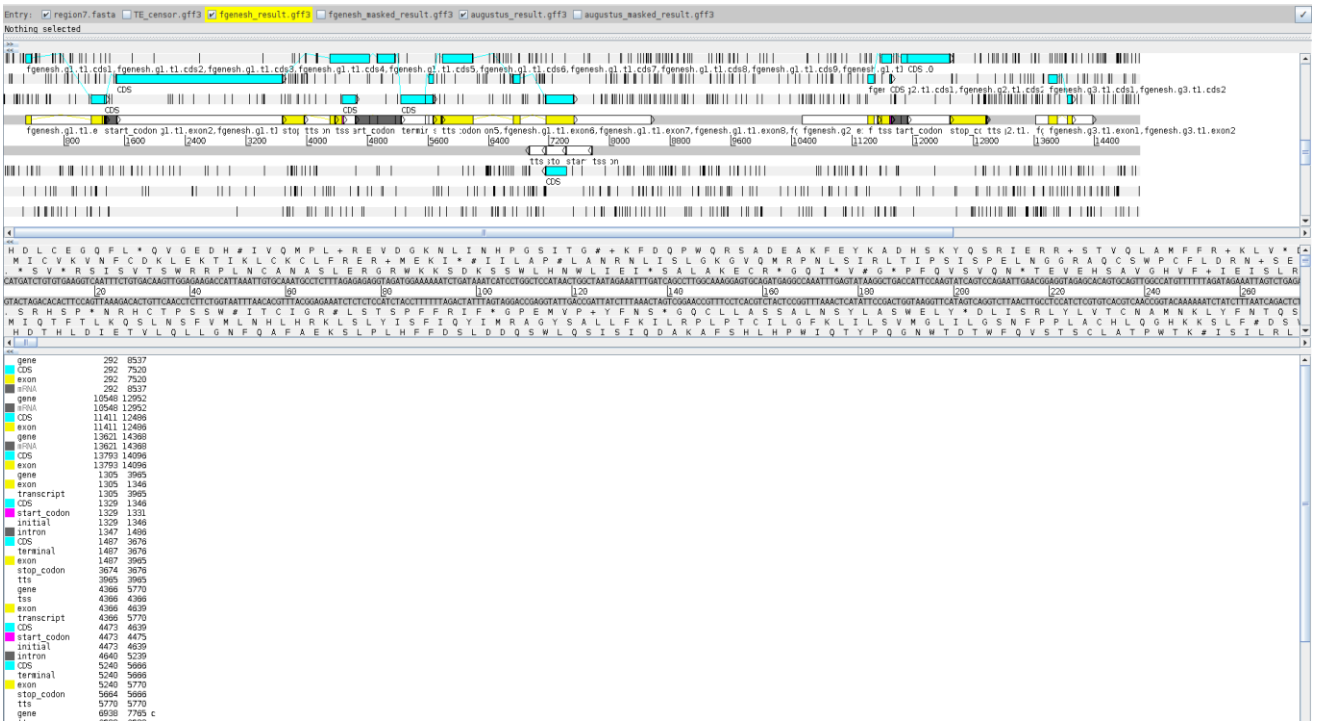


Figure 9. Artemis view of Region 7 showing unmasked Augustus and unmasked FGENESH predictions displayed simultaneously. Four gene models are predicted in total, with several CDS shared between the two predictors, but FGENESH identifies additional CDS segments and alternative exon structures compared to Augustus.

CONCLUSION

The annotation of this wheat genomic region highlights the major difficulties associated with analyzing large, polyploid, and repeat-rich plant genomes. The high abundance of transposable elements and low-complexity sequences can strongly interfere with ab initio gene prediction, leading to the identification of spurious genes and incorrect gene structures if repeat masking and validation are not carefully applied. By integrating structural prediction, repeat annotation, homology-based validation, and genome visualization, it becomes possible to discriminate genuine protein-coding loci from artifacts generated by repetitive DNA. This combined approach not only improves the accuracy of gene models but also provides insight into the organization of genes and transposable elements within the region. Overall, this work emphasizes that reliable genome annotation in Triticum aestivum requires the systematic integration of multiple complementary methods rather than reliance on a single prediction tool.

REFERENCES

1. Australian Center for International Agricultural Research (2019). Retrieved from https://www.aciar.gov.au/media-search/blogs/harvest-knowledge-sequencing-wheat-genome#:~:text=The%20wheat%20genome%20turned%20out,six%20copies%20of%20each%20chromosome.
2. Mario Stanke, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, Burkhard Morgenstern, AUGUSTUS: *ab initio* prediction of alternative transcripts , *Nucleic Acids Research*, Volume 34, Issue suppl_2, 1 July 2006, Pages W435–W439, https://doi.org/10.1093/nar/gkl200
3. Hoff, K. J., & Stanke, M. (2019). Predicting Genes in Single Genomes with AUGUSTUS. *Current protocols in bioinformatics*, *65*(1), e57. https://doi.org/10.1002/cpbi.57
4. https://urgi.versailles.inrae.fr/Data/Transposable-elements/Wheat