De novo annoation of a *Triticum aestivum* genomic region

Module: M1 Génomique structurale

Written by

Christos Mitsakopoulos, Denys Buryi, Masoud Ghanaatian

Table of Contents

Introduction

Our task was to fully annotate an unknown genetic region (region 6) from an unknown *T. aestivum* (Common Wheat) sequenced genome. While ab initio gene prediction software such as AUGUSTUS or FGENESH are built to work with eukaryotic, and even specific plant species (ex. *T. aestivum*), we decided to perform additional anlysis to ensure the validity of the software's predictions. For one, we compared the region itself against popular gene and protein databases, but also looked for a reference genome to which we could perform preliminary analysis with. For this purpose, we chose to use the IWGSC RefSeq v2.1 Common Wheat genome that is the preferred reference genome[1] on National Center for Biotechnology Information (NCBI).

Just like with gene prediction, the primary challenges with successful sequencing and re-assembly of plant genomes, such as that of *T. aestivum*, are defined by elevated genomic ploidy, as well as high proportions of repetitive sequences, transposable elements etc. Specifically, the genome of *T. aestivum* is a hexaploid one, expressed as AABBDD, consisting of the diploid: *T. durrum* tetraploid AABB and *Aegilops Tauschii* derived DD (See Figure 1).

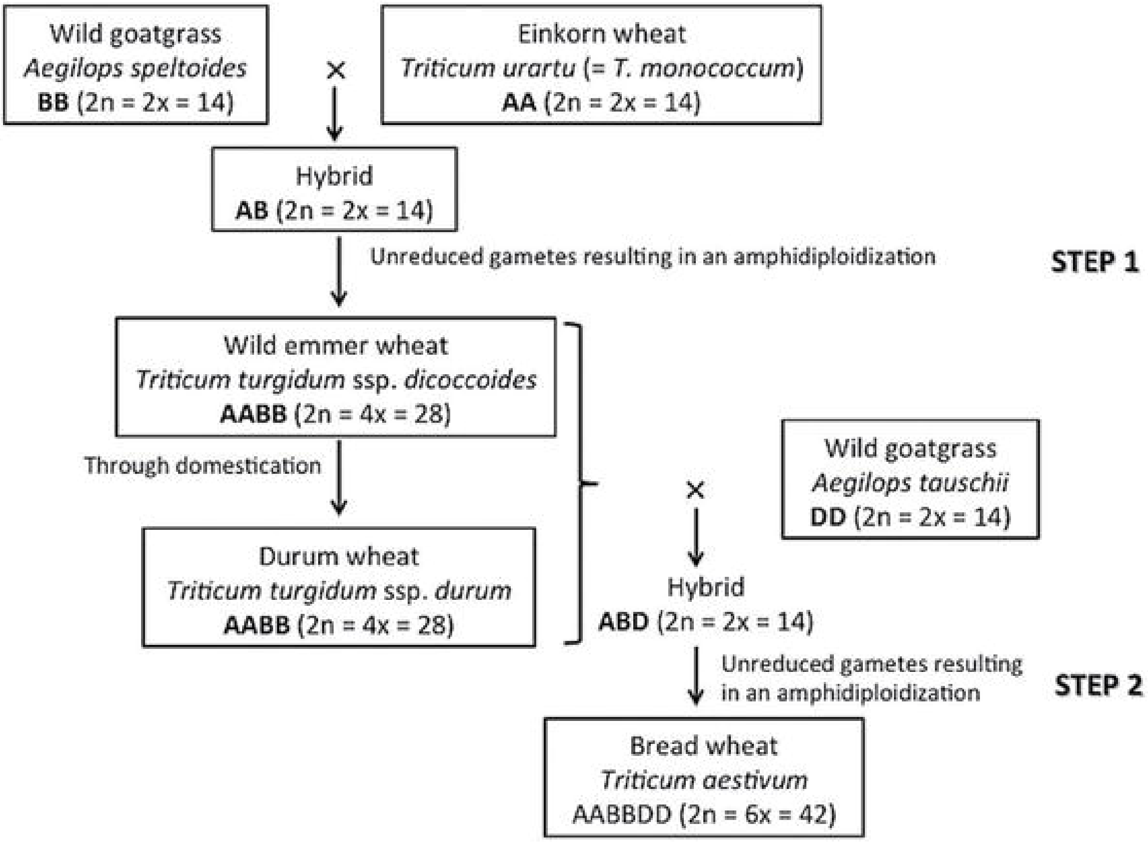


Figure 1: **Figure 1**:*T. aestivum* evolutionary history

Our choice in using IWGSC RefSeq v2.1 as a reference genome for our analysis, is based on the researcher's efforts to refine the original de novo IWGSC RefSeq v1.0 by means of: direct labeling and staining (DSL) (optical mapping) to obtain chromosomes of higher contiguity, as well as filling or bridging gaps in v1.0 scaffolds which were not replaced by DSL data[2]. Importantly, this reference genome contains all three different genome types (ex. AA, BB, DD), in-tact, allowing us to observe in downstream analysis, not just where our unknown region could belong to on the polyploid genome of *T. aestivum*, but also infer how it might have arisen evolutionarily.

Given that the unknown region is 15000 (fifteen-thousand) base pairs(bp) long, with the average size of a gene in the *T.* genus being 3,207 bp, we should expect to see the coding sequences (CDSs) of a limited amount of genes; with regions of low complexity and tandem repeats interspersed throughout, due to the nature of plant genomes[3].

Methods

Note that the initial methodology used the "region6.fasta" without masking, but the analysis was performed again using a RepeatMasker-masked ".fasta", to ensure more reliable results. In an effort to be concise, the results that we obtained in the "unmasked" analysis will be omitted. All files produced in our analysis and relevant annotation images, can be found at our dedicated GitHub repository (`https://github.com/cmitsakopoulos/Region6_annotation`).

RepeatMasker

The HMM based RepeatMasker [4] (version 4.1.7) distribution on GitHub, allowed us to scan the region for low complexity areas(to "mask" them), as well as for transposable elements. The program will return multiple files as output, including the genomic region with masking applied to it (in ".fasta" format), genomic region statistics and "gff3" annotation file of the genomic region if chosen. The distribution worked best when installed locally on MacOS (Ventura 13.0.1), compared to its online tool version on Galaxy[5].

To install it, one must be mindful of the dependencies required by RepeatMasker, which include: Perl (v. 5.40), the Tandem Repeat Finder (TRF[6]) tool, and RMBLAST (A version of CML BLAST that works with RepeatFinder[7]) which we required for our specialised analysis of the genomic region.

To note, RepeatMasker can be installed without RMBLAST, however this tool enables RepeatMasker to use a curated database of Transposable Elements (TEs), to use as reference when searching within the genomic data it is provided. The curated database we used is the TRansposable Elements Platform (TREP) created by the University of Zurich[8]. This database integrates TEs identified and classified within for example, maize or wheat species, but has over the years been updated to include TEs from other species or even genera[8].

RepeatMasker was called in a zsh environment, with the following command:

```
perl ./RepeatMasker \
  -lib /Users/user/Desktop/RepeatMasker/
  Libraries/trep-db_complete_Rel-19.fasta \ #Path to TREP
     database
  -excln \ #Ignore ’N’ used for scaffold connections
  -gff \ #return gff3 with all relevant data
  -gccalc \ #GC calculations for each low compl. region
  -engine rmblast \ #Use RMBLAST dependency
  -poly \ #Apply polymorphism
  /Users/user/Desktop/region6.fasta
```

With the identified TEs at hand, we used the TREP’s ”Search by keyword”[9] function to identify the exact nature of what RepeatMasker had annotated on our genomic region. Once we had identified the TEs in our region, we proceeded with using the masked fasta file that RepeatMasker provides, on the assortment of gene prediction software we will discuss next.

AUGUSTUS

With the masked sequence at hand, we passed it to the “AUGUSTUS” gene prediction software (v.3.4.0) which can be used on the online platform Galaxy [5](v. galaxy2). Predicted genetic coding sequences (CDSs) and their translated protein sequences, were obtained with the following parameters: opting for complete genetic model, with reference organism set to “*T. aestivum*” as is provided in the Augustus “builtin” training set and lastly, softmasking set to 1.

The following parameters were set to true on the AUGUSTUS application on Galaxy:

```
(--protein) #return predicted protein sequences
(--codingseq) #return gene coding sequences
(--introns) #return introns
(--start) #return start codons
(--stop) #return stop codons
(--cds) #return exons
```

FGENESH

FGENSH much like which will be discussed in a subsequent section of the report, served as an auxilliary solution to ab initio prediction of genes and thereafter, annotation of the unknown region. FEGENSH, a Hidden Markov Model (HMM) gene prediction software, was used through its official website offered by 'Softberry'[10]. No parameters for the search can be chosen on the webpage, therefore it should be marked as 'default' parameters.

To ensure reproducibility of this project, we reran FGENESH multiple times and no difference was observed in the prediction score ('1813.750195') and of course, predicted features of our unknown region. This could be owed to the absence of a seed setting and therefore, no randomization of the prediction process in early calculation stages (deterministic algorithm).

geneID

A third option of ab initio prediction of genes, was GeneID of which developers advertise computational efficiency and prediction accuracy[11]. Due to being last updated in 2007[11], the platform is incompatible with "gff3" files, only returning "gff2" on its dedicated web server; therefore be aware of compatibility problems with your desired annotation platform.

Importantly, after encountering problems with annotating the "region6.fasta" in a MacOS (13.0.1 : 22A400) distribution of 'Artemis'[12] with the "gff2", we questioned the source of the "index out of range" error, considering the masked ".fasta" file that was provided to geneID was not corrupted, neither badly formatted. Taking a further look into the gff2 file, the problem was quite obvious:

```
# Acceptors(-) predicted in sequence region6: [0,15000]
region6 geneid_v1.2 Acceptor  15001 15002  0.16 - . #
   NNNNNNNNNNNNNNNNNNNGA
```

However, this small fix uncovered more glaring problems with the "gff2", as the feature keys (ex. "Start") were perceived as invalid by Artemis. Trying to manipulate the names of the keys (ex. "Start" to "Start Codon"), did not seem to solve the problem either; neither is it recommended to convert a gff2 file to a gff3 file[13]. Additionally, other output options included file formats which are proprietary ("geneID format") and cannot be read by either Artemis or 'Jalview'[14], as well as other outdated formats. Therefore, annotations will be omitted from the results section for this program.

If one wants to repeat the analysis process, select the following on the web-server application on the official geneID website[11]:

```
Organism: Triticum aestivum (wheat)
Prediction modes: Normal mode (signal, exon and gene prediction).
DNA strands: Forward and Reverse
Output format: GFF
```

BLAST Tools

Using the Basic Local Alignment Search Tool (BLAST) we performed iterative searches, combining BLAST functionalities to solidify our understanding of the unknown region. Starting with BLASTn, built for nucleotide searches, we investigated the region's alignment against the IWGSC v2.1 reference genome, to obtain the region's possible genomic position, or in other words, the chromosome to which it aligns. In this search we kept most parameters at default, only altering the word size, setting it to 64 in order to avoid the emergence of fragmented homologues in the results (improve homologous hit "contiguity") [15]. In doing so, we could infer the genomic dynamics affecting the unknown region, particularly in how it might have been transferred across or within chromosomes, or from which species it was derived from, during *T. aestivum* evolution (which genome type: AA, BB or DD).

Following this analysis, we wanted to compare genes predicted by automatic tools to data from various databases available online in order to make use of expert annotation and expression data. First, we utilised BLASTn (default settings) using the unknown region as query against the Core_nt (nucleotide) database on NCBI, with the reference organism set to *T. aestivum* of taxid:4565. The returned BLAST ”hits” were intended to be a point of reference to verify the AUGUSTUS predicted CDSs. Unfortunately, upon further examination, the hits we received using this database were dominated by hypothetical genes predicted by Gnomon gene prediction software. As it was not our goal to use predictions of one automatic pipeline to validate predictions of another, we instead concentrated on 2 other databases: Expressed sequence tags database (EST) - which contains libraries of expressed genetic material and UniProtKB/Swiss-Prot (swissprot) protein database, which is expertly annotated and very useful for functional annotation.

First, we used BLASTx with our region as query, against the Swissprot database (BLOSUM80) to identify possible homologs of proteins in our region. We used BLOSUM80 substitution matrix since we were hoping to find highly conserved domains first. Next we used BLASTn with our region as query, against EST database. The results of this search where dominated by expression in the region 1-3780. To sidestep this issue we used the fragment of our region 3781 - 15000 as query in the search with the same parameters, in order to estimate expression of the hypothetical proteins in the second part of our sequence. To further estimate the expression of predicted genes in our region we have performed Blastn against transcriptome shotgun assembly database (TSA). While it is an archive of computationally assembled transcript sequences, it is based on primary data that has been experimentally derived by the same submitter, so it might serve as further evidence of expression. Parameters of the alignment: organism = Triticeae, species specific repeats = T. aestivum, query = full region, rest = default.

Lastly we used the data from the two BLASTn alignments to EST and the alignment to TSA to help estimate the exact coordinates of CDSs of our proteins.

Analysis Results

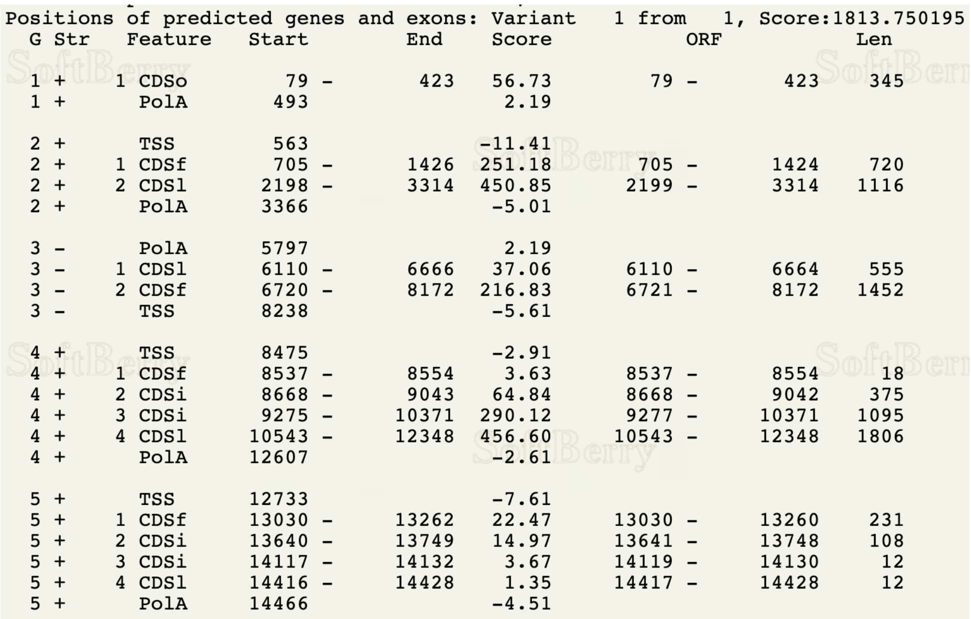
Source Genome

While we do not have access to the originally sequenced T. aestivum genome, using the official reference Common Wheat, IWGSC Refseqv2.1, as classified on NCBI, we investigated the unknown region's position on the reference genome through a BLASTn search primed to identify homologues of high contiguity.

Specifically, we observed that the unknown region has a complete alignment (complete query coverage) with exceptionally high sequence identity (99.99%) to chromosome 4D of the reference. Moreover, we recorded alignments of shorter coverage to the query (˜40% query coverage), but significantly high sequence identity (>90%), to chromosomes 4B and 5A of the reference. With the above results, our first conclusion is that the unknown region is almost certainly located on chromsome 4D in its original sequenced genome; given its alignment with chromosome 4D of IWGSCv2.1, not only has it been inherited from *A. tauschii* (Genome type D) during *T. aestivum's* evolutionary history, considering the negligible nucleotide differences in the IWGSCv2.1 alignment. Moreover, we can assume possible intra-genomic duplication of our region and possibly the genetic code surrounding it, based on the uncharacteristic alignments against chromosomes 4B and 5A of IWGSCv2.1. Serving as more of a preliminary analysis, these conclusions give us the necessary insight to more accurately interpret the rest of our results.

FGENESH

Despite being an auxiliary option for annotating our genomic region, the results of FGENESH are not too far off from AUGUSTUS. Particularly, FGENESH predicted 5 genes on our genomic region, with a cumulative prediction score of "1813.750195" and a total of 13 exons. All information regarding the predicted genes is displayed in the condensed Figure 2.



```
Positions of predicted genes and exons: Variant   1 from    1, Score:1813.750195
 G Str   Feature    Start        End    Score           ORF           Len

 1 +   1 CDSo       79 -       423   56.73         79 -       423   345
 1 +     PolA      493                2.19

 2 +     TSS       563              -11.41
 2 +   1 CDSf      705 -      1426  251.18        705 -      1424   720
 2 +   2 CDSl     2198 -      3314  450.85       2199 -      3314  1116
 2 +     PolA     3366               -5.01

 3 -     PolA     5797                2.19
 3 -   1 CDSl     6110 -      6666   37.06       6110 -      6664   555
 3 -   2 CDSf     6720 -      8172  216.83       6721 -      8172  1452
 3 -     TSS      8238               -5.61

 4 +     TSS      8475               -2.91
 4 +   1 CDSf     8537 -      8554    3.63       8537 -      8554    18
 4 +   2 CDSi     8668 -      9043   64.84       8668 -      9042   375
 4 +   3 CDSi     9275 -     10371  290.12       9277 -     10371  1095
 4 +   4 CDSl    10543 -     12348  456.60      10543 -     12348  1806
 4 +     PolA    12607               -2.61

 5 +     TSS     12733               -7.61
 5 +   1 CDSf    13030 -     13262   22.47      13030 -     13260   231
 5 +   2 CDSi    13640 -     13749   14.97      13641 -     13748   108
 5 +   3 CDSi    14117 -     14132    3.67      14119 -     14130    12
 5 +   4 CDSl    14416 -     14428    1.35      14417 -     14428    12
 5 +     PolA    14466               -4.51
```

Figure 2: *All FGENESH prediction data on our genomic region. Start and Stop sites demonstrate positions on the genomic region. "Pos." stands for position on the genomic region. TSS: ex. TATA-box. CDSf: first Coding Sequence (CDS) in ORF includes start codon. CDSi: internal CDSs in ORF. CDSl: last CDS in ORF includes stop codon. PolyA: PolyA-tail* [16]

Apart from the gene predictions which serve as the forefront of our report, we should note of the identification of 3' untranslatable region (UTR) CDS, these elements promote the Polyadenylation of mRNA, enabling maturation and importantly, assisting in alternative splicing (produce PolyA-tails). One can observe their placement at the 3' end of each predicted gene in Figure 2 (note that gene 3 is on the antisense strand; therefore reverse positioning of PolyA).

Using a perl script to obtain the results of FGENESH from its html file, we could identify the average GC content percentage for each of the five predicted genes: 71.88%, 68.98%, 51.99%, 66.09% and 59.13% respectively.

AUGUSTUS

Following analysis with AUGUSTUS we obtained 4 predicted genes and translated protein sequences, all of varying sizes. On should note that unlike FGENESH, AUGUSTUS did not return any UTR elements (ex. PolyA-tail), both in running it with the masked and unmasked versions of "region6.fasta". Nevertheless, the software designated its predicted proteins with the region's name as prefix and the predicted gene ("g") as suffix: **region6.g1**, **region6.g2**, **region6.g3**, **region6.g4** (Refer to Table 1 for relevant information).

Table 1	**region6.g1**	**region6.g2**	**region6.g3**	**region6.g4**
Prediction Score	0.62	0.93	0.98	0.87
Total Size (n)	2609	1526	3650	245
Start, Stop Pos.	705-3314	6646-8172	8698-12348	13030-13275
Introns (GC%)	1 (41.42)	0	0	0
Start, Stop Pos.	1271-2197	NA	NA	NA
Exons (GC%)	2 (70.11/53.83))	1 (61.27)	1 (56.5)	1
Start, Stop Pos.	705-1270 / 2198-3314	6646-8172	8698-12348	13030-13275
Strand	sense (+)	antisense (-)	sense (+)	sense (+)

Table 1: *Depicts basic information regarding the genetic makeup of the predicted genes. "Pos." stands for position on the genomic region. "Start and Stop" positions of predicted features on the unknown region. Prediction scores calculated for the predicted genes by AUGUSTUS, range from 0 to 1. "Strand" refers to the strand which is transcribed. GC content was calculated using the Artemis annotation software.*

An "html" of the masked region annotated with AUGUSTUS results can be found on our GitHub in the relevant "AUGUSTUS" folder (https://github.com/cmitsakopoulos/Region6_annotation)(Created with Jalview).

GeneID

Due to outdated file formats and 3rd party software incompatibilities with GeneID, the results section will be comparatively limited. To start with, GeneID returned two predicted genes, which it named "Gene 1", "Gene 2". We managed to source the following information from the "gff2", regarding these two predicted genes:

```
# Column 4: start / Column 5: end / Column 6: prediction score
# Gene 1 (Forward). 2 exons. 561 aa. Score = 136.29
region6 geneid_v1.2 First Exon  705 1270  53.20 + 0 region6_1
region6 geneid_v1.2 Terminal  2198  3314  83.09 + 1 region6_1
# Gene 2 (Forward). 5 exons. 1067 aa. Score = 132.23
region6 geneid_v1.2 First Exon  7148  7231   0.56 + 0 region6_2
region6 geneid_v1.2 Internal Exon 8519  8554   0.36 + 0 region6_2
region6 geneid_v1.2 Internal Exon 8668  8801   4.99 + 0 region6_2
region6 geneid_v1.2 Internal Exon 9231  10873 60.84 + 1 region6_2
region6 geneid_v1.2 Terminal  11045 12348 65.48 + 2 region6_2
```

RepeatMasker

The program identified that our region has a GC content of 53.95% and was able to mask 9.42% (or 1413bp) for low complexity. Following RMBLAST, we observe some evidence of simple repeats which comprise 2.85% of the region. Alongside the simple repeats, a limited number of TEs were identified, with really high alignment scores (SW-Positive penalties). Refer to Table 3 which illustrates the aforementioned SW alignments.

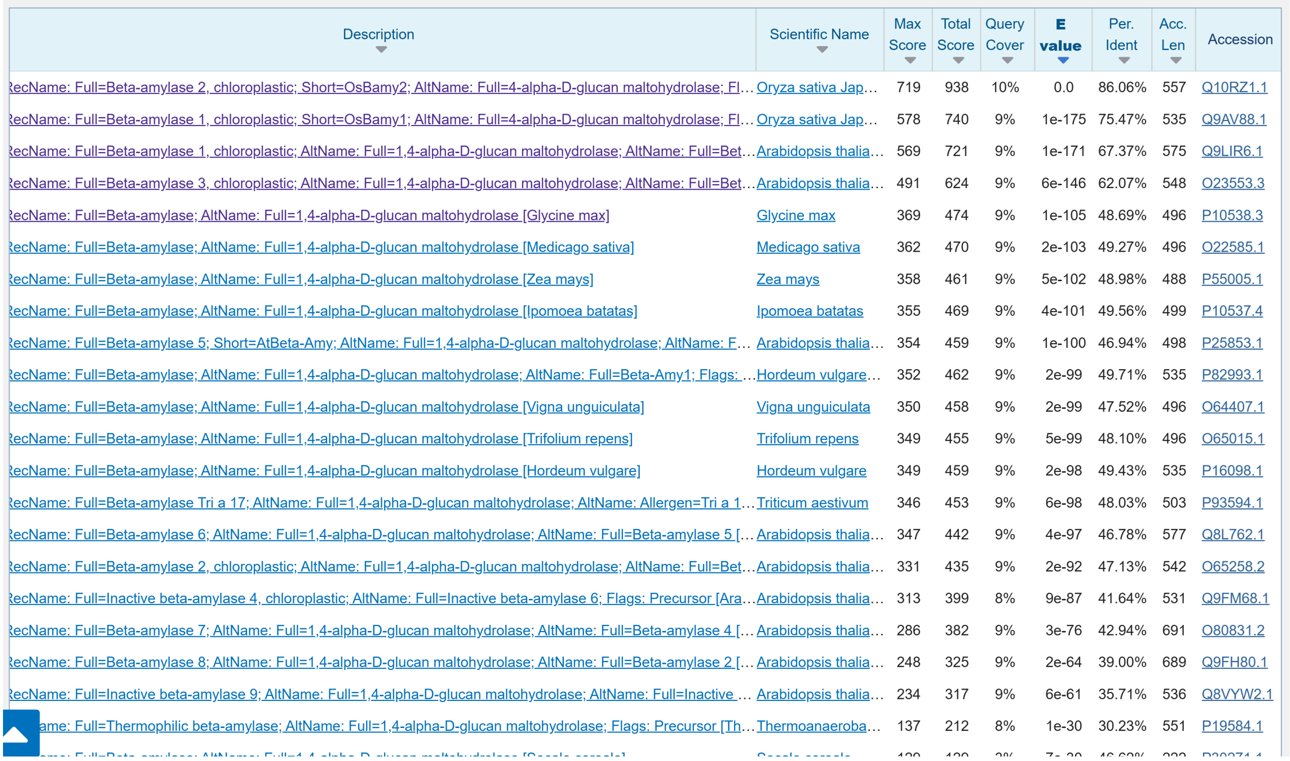
SW score			Query					Matching
Score	perc (% div/del/ins)		Position					
			Sequence	Begin	End	(Left)		Repeat
15	0.0	0.0/0.0	region6	683	701	(14300)	+	(CGT)n
321	19.8	1.1/1.1	region6	1998	2089	(12912)	C	TE ”fragment?” *T. aestivum*
391	14.1	0.0/1.3	region6	2004	2082	(12919)	C	Retrotransposon, non-LTR (SINE)
251	20.6	1.2/6.3	region6	**3648**	**3730**	(11271)	C	TE: Complete Element
337	14.3	1.6/0.0	region6	3668	3730	(11271)	C	Stowaway MITE; complete element
279	25.7	1.3/1.3	region6	**3731**	**3805**	(11196)	C	**Stowaway-MITE**
614	24.6	5.5/4.3	region6	3984	3999	(11002)	+	TE: Complete element
40	0.0	0.0/0.0	region6	4000	4033	(10968)	+	(CA)n
614	24.6	5.5/4.3	region6	4034	4271	(10730)	+	TE: Complete element
27	37.1	1.4/1.4	region6	9048	9194	(5807)	+	(GCC)n
16	21.0	4.5/4.5	region6	10399	10464	(4537)	+	(GCG)n
11	21.9	3.9/3.9	region6	10490	10540	(4461)	+	(TGGCGA)n
401	27.0	0.0/1.6	region6	12432	12559	(2442)	+	DNA-transposon, TIR, CACTA
12	10.0	0.0/0.0	region6	13937	13958	(1043)	+	(GCAT)n
330	30.3	7.0/0.6	region6	14152	14307	(694)	+	DNA-transposon, TIR, CACTA
415	23.9	0.9/0.0	region6	14153	14265	(736)	C	DNA-transposon, TIR, CACTA
15	30.9	0.0/6.0	region6	14485	14572	(429)	+	(AAAAAGA)n
236	26.2	8.6/5.4	region6	**14713**	**14910**	(91)	+	**Copia LTR-Retrotransposon**
239	28.7	5.3/2.2	region6	**14737**	**14868**	(133)	C	**SINE-retrotransposon**

Table 2: *The results displayed in the table can be seen in the ”region6.fasta.tbl” file attached to the GitHub repository[17]. The Smith-Waterman(SW) score is obtained from a scoring matrix during local alignment of RepeatMasker's RMBLAST. ”Div.”, ”Del.” and ”Ins.” amount to the percentage of alignment: divergence (mismatching), deletion (gaps) and insertion (gap due to insertion) respectively. Finally, the Repeat column contains the type of Repeat, obtained from the TREP database with the RMBLAST returned IDs of SW alignments. ”C” stands for complementary strand; complementary to the sense(+) strand. (See more[18])*

Reading from Table 2, one can see that despite using a curated database of *T. aestivum* TEs, our RepeatMasker results did not meet expectations. While we have strong evidence for simple repeats (7), the RMBLAST TE hits exhibit high SW alignment scores (Positive penalties). Considering literature indicating that ˜80% of the *T. aestivum* genome comprises TEs[19] and acknowledging the propensity of TEs to mutate, we continue to analyse them regardless. With a SW score threshold of 300, we get the best (lowest) scoring TEs. The lowest scoring TE, is a *Copia* family Long Terminal Repeat (LTR)-Retrotransposon from the *Hordeum vulgare* species, with an alignment score of 236. The second best scoring TE, is a Short Interspersed Nuclear Element(SINE) retrotransposon from the *Triticum monococcum* species, with a score of 239. The next two TEs that are below the threshold are both derived from *T. aestivum* and are: a Stowaway Miniature Inverted-Repeat Element (MITE) and a complete element (capable of independent transposition). Unfortunately, further documentation on these specific transposable elements is very limited and research on Google Scholar did not help us identify the family of these TEs, other than the *Copia* one.

BLASTx and BLASTn

Our first BLASTx run, as described in Methods produced an intriguing set of alignments, which are displayed in Figure 4 below.



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
RecName: Full=Beta-amylase 2, chloroplastic; Short=OsBamy2; AltName: Full=4-alpha-D-glucan maltohydrolase; Fl…	Oryza sativa Jap…	719	938	10%	0.0	86.06%	557	Q10RZ1.1
RecName: Full=Beta-amylase 1, chloroplastic; Short=OsBamy1; AltName: Full=4-alpha-D-glucan maltohydrolase; Fl…	Oryza sativa Jap…	578	740	9%	1e-175	75.47%	535	Q9AV88.1
RecName: Full=Beta-amylase 1, chloroplastic; AltName: Full=1,4-alpha-D-glucan maltohydrolase; AltName: Full=Bet…	Arabidopsis thalia…	569	721	9%	1e-171	67.37%	575	Q9LIR6.1
RecName: Full=Beta-amylase 3, chloroplastic; AltName: Full=1,4-alpha-D-glucan maltohydrolase; AltName: Full=Bet…	Arabidopsis thalia…	491	624	9%	6e-146	62.07%	548	O23553.3
RecName: Full=Beta-amylase; AltName: Full=1,4-alpha-D-glucan maltohydrolase [Glycine max]	Glycine max	369	474	9%	1e-105	48.69%	496	P10538.3
RecName: Full=Beta-amylase; AltName: Full=1,4-alpha-D-glucan maltohydrolase [Medicago sativa]	Medicago sativa	362	470	9%	2e-103	49.27%	496	O22585.1
RecName: Full=Beta-amylase; AltName: Full=1,4-alpha-D-glucan maltohydrolase [Zea mays]	Zea mays	358	461	9%	5e-102	48.98%	488	P55005.1
RecName: Full=Beta-amylase; AltName: Full=1,4-alpha-D-glucan maltohydrolase [Ipomoea batatas]	Ipomoea batatas	355	469	9%	4e-101	49.56%	499	P10537.4
RecName: Full=Beta-amylase 5; Short=AtBeta-Amy; AltName: Full=1,4-alpha-D-glucan maltohydrolase; AltName: F…	Arabidopsis thalia…	354	459	9%	1e-100	46.94%	498	P25853.1
RecName: Full=Beta-amylase; AltName: Full=1,4-alpha-D-glucan maltohydrolase; AltName: Full=Beta-Amy1; Flags: …	Hordeum vulgare…	352	462	9%	2e-99	49.71%	535	P82993.1
RecName: Full=Beta-amylase; AltName: Full=1,4-alpha-D-glucan maltohydrolase [Vigna unguiculata]	Vigna unguiculata	350	458	9%	2e-99	47.52%	496	O64407.1
RecName: Full=Beta-amylase; AltName: Full=1,4-alpha-D-glucan maltohydrolase [Trifolium repens]	Trifolium repens	349	455	9%	5e-99	48.10%	496	O65015.1
RecName: Full=Beta-amylase; AltName: Full=1,4-alpha-D-glucan maltohydrolase [Hordeum vulgare]	Hordeum vulgare	349	459	9%	2e-98	49.43%	535	P16098.1
RecName: Full=Beta-amylase Tri a 17; AltName: Full=1,4-alpha-D-glucan maltohydrolase; AltName: Allergen=Tri a 1…	Triticum aestivum	346	453	9%	6e-98	48.03%	503	P93594.1
RecName: Full=Beta-amylase 6; AltName: Full=1,4-alpha-D-glucan maltohydrolase; AltName: Full=Beta-amylase 5 […	Arabidopsis thalia…	347	442	9%	4e-97	46.78%	577	Q8L762.1
RecName: Full=Beta-amylase 2, chloroplastic; AltName: Full=1,4-alpha-D-glucan maltohydrolase; AltName: Full=Bet…	Arabidopsis thalia…	331	435	9%	2e-92	47.13%	542	O65258.2
RecName: Full=Inactive beta-amylase 4, chloroplastic; AltName: Full=Inactive beta-amylase 6; Flags: Precursor [Ara…	Arabidopsis thalia…	313	399	8%	9e-87	41.64%	531	Q9FM68.1
RecName: Full=Beta-amylase 7; AltName: Full=1,4-alpha-D-glucan maltohydrolase; AltName: Full=Beta-amylase 4 […	Arabidopsis thalia…	286	382	9%	3e-76	42.94%	691	O80831.2
RecName: Full=Beta-amylase 8; AltName: Full=1,4-alpha-D-glucan maltohydrolase; AltName: Full=Beta-amylase 2 […	Arabidopsis thalia…	248	325	9%	2e-64	39.00%	689	Q9FH80.1
RecName: Full=Inactive beta-amylase 9; AltName: Full=1,4-alpha-D-glucan maltohydrolase; AltName: Full=Inactive …	Arabidopsis thalia…	234	317	9%	6e-61	35.71%	536	Q8VYW2.1
ame: Full=Thermophilic beta-amylase; AltName: Full=1,4-alpha-D-glucan maltohydrolase; Flags: Precursor [Th…	Thermoanaeroba…	137	212	8%	1e-30	30.23%	551	P19584.1
ame: Full=Beta-amylase; AltName: Full=1,4-alpha-D-glucan maltohydrolase [Secale cereale]	Secale cereale	129	129	3%	7e-30	46.62%	222	P30271.1

Figure 3: *BLASTx top results, region 6 = query, UniProtKB/Swiss-Prot database, BLOSUM80*

Especially relevant are the first 4 results, 2 different chloroplastic beta-amylases in 2 organisms: *Oryza Sativa* and *Arabidopsis thaliana*. Each of these alignments consist of 2 ranges with approximately the same coordinates, each with very high % identity (id), very low e-values, and very good alignments. For the rest of the beta-amylase results, % id decreases significantly, but one could argue that these are homologous proteins.

In a different part of our region we have 4 alignments to 4 different E3 ubiquitin-protein ligases in *A. thaliana*. All of them have relatively low % identity and only 2 are substantial: them being relatively long, having 31 and 35 % identity, as well as a confirmed and annotated protein. The WAVH2 protein is annotated as ”Probable”, and alignment to EDA40 is very short, with a very low e-value. (Figure 4)

The rest of the alignments are of poorer quality: very low e-values, hits on more distantly related species, and therefore less clear possibility of homology, increasing our difficulties in interpreting the results.

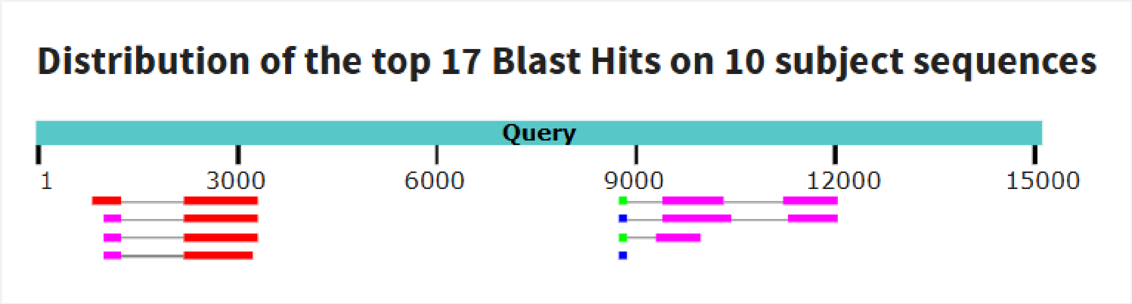

Distribution of the top 17 Blast Hits on 10 subject sequences
Query
1
3000
6000
9000
12000
15000

Figure 4: *BLASTx results distribution, region 6 = query, UniProtKB/Swiss-Prot database, BLOSUM80. Distribution given for the 4 top alignments and the 4 E3 ubiquitin-protein ligase alignments*

In order to complement these findings we will combine the results of BLASTn (of our region) against the EST database, to observe if the genes coding for these proteins seem to be expressed in *T. aestivum*. First alignment was performed with default parameters except for species-specific repeats filter (*T. aestivum*). The top 100 alignments have very high % id (84+, most 95+), most of them align to mRNA in *T. aestivum* and all of them align to the gene likely coding for amylase, and from their distribution we can estimate that the two gene exons are located between ˜ 450 to ˜ 3620 bp.

To explore the expression data on the rest of the sequence we have performed BLASTn with the same parameters for the sequence starting at 3781 bp to the end:

```
NB! for the coordinates mentioned below, the absolute coordinate
   in the region is = x + 3780.
```

We have performed alignment vs EST database with organism = *T. aestivum*, and species specific repeats set to the same. We have only obtained 2 small alignments (300 and 400 bp) in the region. So we have decided to expand our search to the tribe *Triticeae* (same parameters + organism = *Triticeae*). We have obtained 41 alignments all of which are relatively short (max 404 bp), but possessing high % identity. Most of the hits align to *H. vulgare* and their distribution is quite peculiar. (Figure 5)

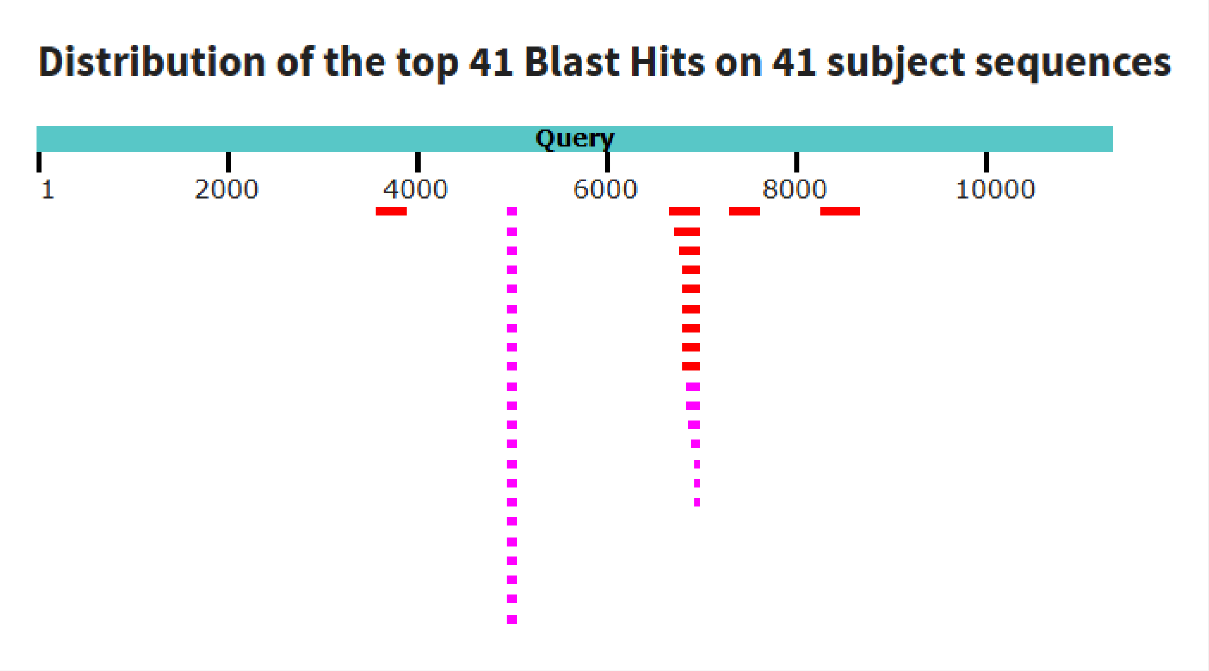
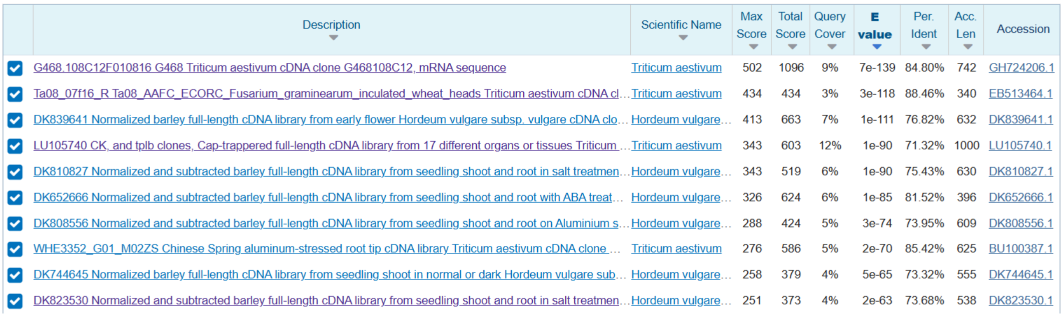


Figure 5: *BLASTn results distribution, region(3781+) = query, EST database, limited to: Triticeae. The first and last alignments are to T. aestivum, and were found in the search mentioned above.*

The set of alignments forming a second "column" ending at ˜ 7000 bp is explained by the search below, it is likely a piece of the gene, of which the other pieces are located downstream. On the other hand, the set of alignments forming a "column" around 5000 bp is harder to interpret. All of them are to *H. vulgare*, all with relatively low e-value and quite short. All but one start at 4911 bp and all end at 5018 bp. Each fragment on their own would likely to have been an accidental hit. All together they imply the possible duplication of fragments. But considering the fragments come from 2 different studies at different locales, it introduces a small likelihood it's either an exon of a downstream gene that has been gained in barley, or an exon that has been lost in wheat (less likely since no hits to other *Triticeae*). It is also possible there exists a rarely expressed in wheat splice variant that includes this fragment as exon. Continuing to explore the expression of genes in this region, we performed BLASTn with the same parameters as in a previous search, except changing "optimizing for" from "Highly similar sequences" to "Somewhat similar sequences". Results of this search are shown in Figure 6 and Figure 7 and discussed below.

Within the top 10, are 4 alignments to mRNA in *T. aestivum* and 6 to mRNA in *H. vulgare*. All of these have high % identity (71+), with relatively low e-values. These results hint at the possible location of another gene between ˜ 6450 bp and 8584 bp. Curiously, for 7 of these alignments the two potential exons end at the position 7072 and 8584 respectively.



	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	G468.108C12F010816 G468 Triticum aestivum cDNA clone G468108C12, mRNA sequence	Triticum aestivum	502	1096	9%	7e-139	84.80%	742	GH724206.1
☑	Ta08_07f16_R Ta08_AAFC_ECORC_Fusarium_graminearum_inculated_wheat_heads Triticum aestivum cDNA cl…	Triticum aestivum	434	434	3%	3e-118	88.46%	340	EB513464.1
☑	DK839641 Normalized barley full-length cDNA library from early flower Hordeum vulgare subsp. vulgare cDNA clo…	Hordeum vulgare…	413	663	7%	1e-111	76.82%	632	DK839641.1
☑	LU105740 CK, and tplb clones, Cap-trappered full-length cDNA library from 17 different organs or tissues Triticum …	Triticum aestivum	343	603	12%	1e-90	71.32%	1000	LU105740.1
☑	DK810827 Normalized and subtracted barley full-length cDNA library from seedling shoot and root in salt treatmen…	Hordeum vulgare…	343	519	6%	1e-90	75.43%	630	DK810827.1
☑	DK652666 Normalized and subtracted barley full-length cDNA library from seedling shoot and root with ABA treat…	Hordeum vulgare…	326	624	6%	1e-85	81.52%	396	DK652666.1
☑	DK808556 Normalized and subtracted barley full-length cDNA library from seedling shoot and root on Aluminium s…	Hordeum vulgare…	288	424	5%	3e-74	73.95%	609	DK808556.1
☑	WHE3352_G01_M02ZS Chinese Spring aluminum-stressed root tip cDNA library Triticum aestivum cDNA clone …	Triticum aestivum	276	586	5%	2e-70	85.42%	625	BU100387.1
☑	DK744645 Normalized barley full-length cDNA library from seedling shoot in normal or dark Hordeum vulgare sub…	Hordeum vulgare…	258	379	4%	5e-65	73.32%	555	DK744645.1
☑	DK823530 Normalized and subtracted barley full-length cDNA library from seedling shoot and root in salt treatmen…	Hordeum vulgare…	251	373	4%	2e-63	73.68%	538	DK823530.1

Figure 6: *BLASTn top 10 results , region(3781+) = query, EST database, limited to: Triticeae, optimize for: somewhat similar*

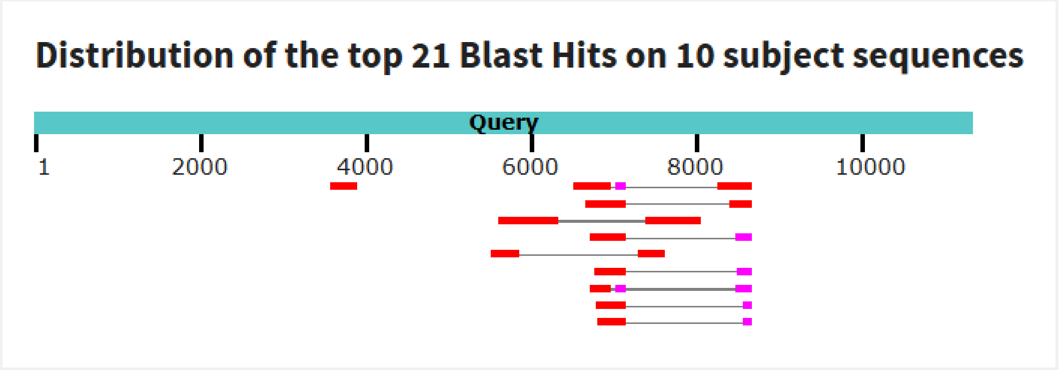


Figure 7: *BLASTn top 10 results distribution, region(3781+) = query, EST database, limited to: Triticeae, optimize for: somewhat similar.*

To further estimate the expression of genes in our region, we performed BLASTn to TSA database (described in Methods). All top alignments were to RNA sequences in T. aestivum, with a very high % identity, and the top 20 alignments were especially good (long, rare gaps, rare low complexity regions). These top 20 results have two or three ranges, but are quite consistent (Table 3). For example, for hypothetical exon 1, all alignments that are present at those coordinates are within a few bp of the coordinate 460 (n=9) or 903 (n = 8) at the start and all except one end at 1270 bp. For the hypothetical exon 2, there are two groups in these results: alignments with the exon being split into two ranges or the exon being contained in one range. Most alignments present in this region (n=17) start at coordinate 2195 (n=15) or within 3 bases of it (n=2). The end of this hypothetical exon 2 is less clear, but there is still some consensus, most end coordinates are between 3521-3727 (n=17). Considering that these sequences are based on mRNA, subject to post-transcription modification and quick deterioration, we cannot expect to see a perfect consensus. With this additional evidence, it is significantly more likely that a protein is expressed at these coordinates (probably a chloroplastic beta-amylase).

Alignment	R2: start	R2: end	R1: start	R1: end	R3: start	R3: end
1	823	1270	2195	4102	-	-
2	450	1270	2195	3642	-	-
3	466	1270	2195	3607	-	-
4	903	1270	2195	3594	-	-
5	903	1270	2195	3594	-	-
6	459	1270	2198	3590	-	-
7	453	1270	2195	3659	-	-
8	464	1270	2195	3627	-	-
9	457	1270	2195	3393	-	-
10*	466	1566	-	-	-	-
11	887	1270	2195	3321	3362	3521
12	923	1270	2195	3321	3362	3721
13	-	-	2493	3704	-	-
14	903	1270	2195	3321	3362	3680
15	903	1270	2195	3321	3362	3680
16*	-	-	2198	3321	3362	3623
17	465	1270	2195	3321	3362	3603
18	829	1270	2195	3321	3362	3727
19	488	1270	2195	3321	3362	3612
20	-	-	2493	3704	-	-

Table 3: *Coordinates of the top 20 hits of BLASTn alignment of our region against TSA database. R1, R2, R3 - range 1, 2, 3 - preserved automatic labelling by BLAST algorithm. Columns presented in order, sorted from left to right by* ***coordinate on the query****. *For these alignments, the coordinates were moved to a different column despite automatic labeling (BLASTn parameters: query = region, results limited to = Triticeae, species specific repeats filter = T. aestivum, optimize for highly similar sequences)*

Estimating protein coordinates - Beta-amylase

Taking into account the evidences available to us at this point we can start estimating the real coordinates of a protein which is most likely a cloroplastic beta-amylase:

1. FGENESH predicted a gene with TSS at 563 bp, exon 1 between 705 - 1426 bp, exon 2 between 2198-3314, end of PolA at 3366 bp
2. AUGUSTUS predicted: exon 1: 705-1270 bp, exon 2: 2198-3314 bp
3. BLASTn vs EST: fragments region 1: start ˜ 450-470 bp, end ˜ 1220-1270 bp; fragments region 2: start 2195 bp (n=5), end ˜ 3621-3630 (n = 17)
4. BLASTn vs TSA: fragments region 1: start ˜ 460 or 903 bp, end 1270 bp; fragments region 1: start 2195 bp, end 3521-3727 bp.

From this information we can reason that part of the mRNA fragments starting at around 903 bp are likely mature mRNA that was shortened post transcription, and the mRNA fragments starting at ˜ 460 bp are likely from pre-mRNA that still includes TSS. This aligns well with software prediction. Looking at the information about the end of exon 1, AUGUSTUS prediction aligns well with the end of majority of mRNA fragments, so we could tentatively conclude that the actual coordinates of gene 1 are:

- 5‘ UTR: 460-705(low p./alternative 903) bp;
- CDS1: 705 (low p./alternative 903) - 1270 bp;
- intron: 1270 - 2195 bp;
- CDS2: 2195 - 3314 bp;
- 3’ UTR: 3314 - 3630 bp

Estimating protein coordinates - E3 ubiquitin-ligase

1. FGENESH predicted a gene with TSS at 8475 bp, exon 1 between 8537-8554 bp, exon 2 between 8668-9043, exon 3 between 9275-10371 exon 4 between 10543-12348 end of PolA at 12607 bp
2. AUGUSTUS predicted exons: CDS1: 8698-12348 bp,
3. BLASTn vs TSA: exon 1: 8691 (n=15) - 8798 (n = 18) exon 2: 9251 (n=26) - 10685 (n=15) or 10700 (n=20) exon 3: 10991 (n=25) - 12364 (n=30)
4. BLASTn vs EST: 2 types of mRNA fragments in this region: fragment 1: between 8691 - 8798 bp (n=8), fragment 2: from ˜ 10419 to 10740 bp (n=14)

Here we need to take a few pieces of information into account. AUGUSTUS predicts just one long exon while FGENESH offers 4 different exons. The EST database does not seem to contain evidence of a full protein being expressed at these coordinates in T. aestivum. Although, the TSA database returns hits that are very consistent with regards to their coordinates. On this information alone we could infer possible duplication of fragments, but these assembled sequences come from different studies at different locations. One other explanation could be a particularity within an assembly algorithm; although for these assemblies authors used at least 2 different ones (scallop 0.10.2 and trinity 2.4.x). One last possibility is how the assembly algorithm interacts with the reference genome, possibly forcing an mRNA that would correspond to a gene predicted on one of the other chromosomes. In summation, we propose that this gene is a homeolog of E3 ubiquitin-ligase with the expressed version of it being located on another chromosome or in another genome. An alternative explanation could be that we're dealing with a pseudo-gene that was disrupted by a transposable element.

Estimating protein coordinates - region6.g2

In addition to the above, we found evidence of expression and software prediction intersecting in one more locus.

1. FGENESH predicted a gene on a "-" strand with TSS at 8238 bp, exon 2 between 6110-6666 bp, exon 1 between 6720-8172, end of PolA at 5797 bp
2. AUGUSTUS predicted exons: CDS1: 6646-8172 bp,
3. BLASTn vs TSA produced 12 alignments to "-" strand: exon 2: 6614 (n=5) - 6820 (n=10) bp exon 1: start - 7011 (n=7) or 7014 (n=3), end 8122 (n = 7) or end 8125 (n=5)
4. BLASTn vs EST did not produce any alignments in this region.

We see an overlap of FGENESH prediction with BLASTn on TSA alignments, although it isn't very close. We might be seeing a transposable element, but RepeatMasker did not identify one in this region (See Table 2). An alternative explanation could be that this sequence is a component of another pseudogene, or another homeologous gene. However, the information we have available is insufficient to draw a conclusion from; warranting further investigation in the future.

Conclusion

Throughout our analysis we integrated multiple methods aimed at ab initio gene prediction and functional annotation. After cross referencing BLAST alignments and the results of gene prediction software, we have identified a gene coding for chloroplastic beta-amylase in T. aestivum. Additionally, after compiling results from gene prediction software and multiple gene expression datasets, we conclude that an apparent E3 ubiquitin-ligase coding gene is a likely homeolog or pseudogene within our region. To end, a region appears to be expressed in some T. aestivum plants, of which we were unable to draw a tangible conclusion on its origin and function.

As required by the project descriptor, a final annotation of our region is depicted below. The figures show what we suspect our confirmed region 6 genes look like; see Figure 9 and Figure 10

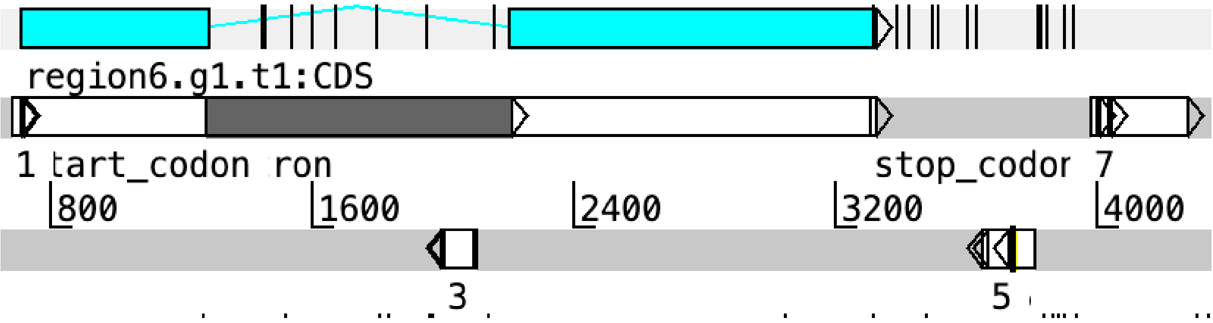


Figure 8: Generated in Artemis using both the RepeatMasker and AUGUSTUS gff files (FEGENESH identical to AUGUSTUS), the screenshot contains the regions at which we have identified possible evidence of a Beta-Amylase. CDS in blue, Entire predicted gene region in white, representation is on the sense strand.

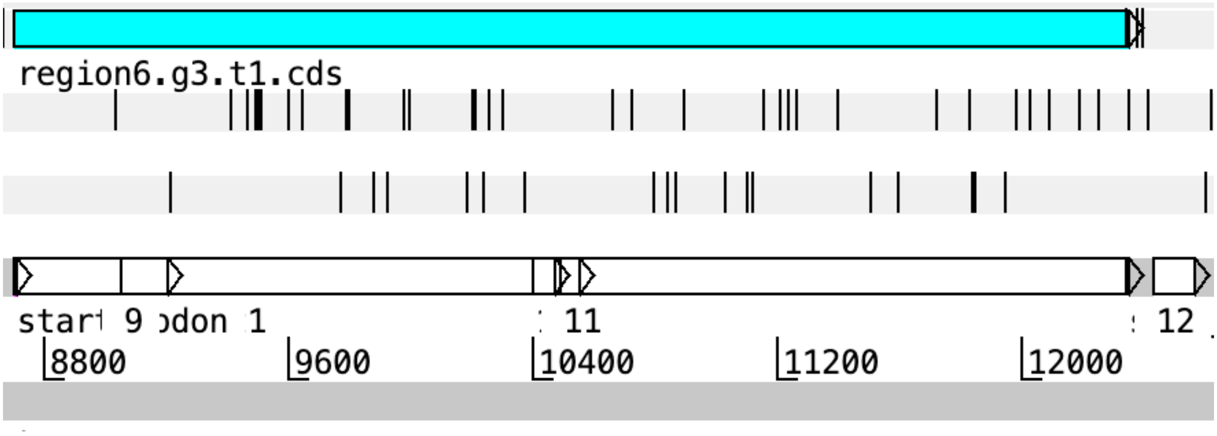


Figure 9: Generated in Artemis using both the RepeatMasker and AUGUSTUS gff files (FEGENESH identical to AUGUSTUS), the screenshot contains the regions at which we have identified possible evidence of an E3-Ubiquitin Ligase. CDS in blue, Entire predicted gene region in white, arrows pointing right are TEs (mostly simple repeats); representation is on the sense strand.

References

[1] Davis University of California. *Triticum aestivum genome assembly IWGSC CS RefSeq v2.1*. 2024. URL: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_018294505.1/.

[2] T. Zhu et al. "Optical Maps Refine the Bread Wheat *Triticum aestivum* cv. Chinese Spring Genome Assembly". In: *The Plant Journal* 107.1 (2021), pp. 303–314. DOI: 10.1111/tpj.15220. URL: https://doi.org/10.1111/tpj.15220.

[3] Hong-Qing Ling et al. "Draft genome of the wheat A-genome progenitor Triticum urartu". In: *Nature* 496.7443 (Apr. 2013), pp. 87–90. DOI: https://doi.org/10.1038/nature11997. URL: https://www.nature.com/articles/nature11997.

[4] Institute for Systems Biology. *RepeatMasker Home Page*. Nov. 2024. URL: https://www.repeatmasker.org/.

[5] IFB NNCR Cluster Task force. *Galaxy — France*. URL: https://usegalaxy.fr/.

[6] G Benson. “Tandem repeats finder: a program to analyze DNA sequences”. In: *Nucleic acids research* 27.2 (1999), pp. 573–80. DOI: https://doi.org/10.1093/nar/27.2.573. URL: https://www.ncbi.nlm.nih.gov/pubmed/9862982.

[7] Institute for Systems Biology. *RMBlast Download Page*. 2025. URL: https://www.repeatmasker.org/rmblast/.

[8] Edith Schlagenhauf. *UZH - IPMB - TREP Database - Welcome to the TRansposable Elements Platform*. 2019. URL: https://trep-db.uzh.ch/index.php.

[9] Edith Schlagenhauf. *UZH - IPMB - TREP Database - TEs by Keyword*. 2022. URL: https://trep-db.uzh.ch/searchTREP.php.

[10] Softberry. *FGENESH - HMM-based gene structure prediction*. 2016. URL: http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind.

[11] Francisco Camara. *geneid homepage*. 2007. URL: https://genome.crg.es/software/geneid/.

[12] T. Carver et al. “Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data”. In: *Bioinformatics* 28.4 (Dec. 2011), pp. 464–469.

[13] GMOD. *GMOD*. 2017. URL: https://gmod.org/wiki/GFF2#Converting_GFF2_to_GFF3.

[14] A. M. Waterhouse et al. “Jalview Version 2–a multiple sequence alignment editor and analysis workbench”. In: *Bioinformatics* 25.9 (Jan. 2009), pp. 1189–1191. DOI: https://doi.org/10.1093/bioinformatics/btp033.

[15] M. Scholz. *Metagenomics - BLAST Word-Size*. Apr. 2023. URL: https://www.metagenomics.wiki/tools/blast/default-word-size (visited on 01/09/2025).

[16] Softberry. *SoftBerry - FGENESH HELP*. 2025. URL: http://www.softberry.com/berry.phtml?topic=fgenesh&group=help&subgroup=gfind.

[17] cmitsakopoulos. *GitHub - cmitsakopoulos/Region6*$_a$*nnotation : Arepositorypurposedforstoringthe* 2025. URL: https://github.com/cmitsakopoulos/Region6_annotation.

[18] AnimalGenome.org. *RepeatMasker Documentation*. 2005. URL: https://www.animalgenome.org/bioinfo/resources/manuals/RepeatMasker.html.

[19] Inbar Bariah, Danielle Keidar-Friedman, and Khalil Kashkush. “Where the Wild Things Are: Transposable Elements as Drivers of Structural and Functional Variations in the Wheat Genome”. In: *Frontiers in Plant Science* 11 (Sept. 2020). DOI: https://doi.org/10.3389/fpls.2020.585515.