MRes Bioinformatics CW 2: Genome Sequence Analysis Assignment

Jude Popham: 01913327

Words: 999 (not including underlined headings)

Task 1:

File 1 (.tbl)

This table summarises the characteristics, classification, and % distribution of repeat content in the query sequence (4 interspersed repeat classes and other repetitive or low-complexity elements). It also details the query sequence's total length, GC content and masked bases.

File 2 (.log)

This is a status message from running RepeatMaster's cross_match. Terms (words) are used to index repeats within the DNA sequence. Pruning refines the index to subset relevant terms.

File 3 (.out)

This file lists repetitive elements found in the query DNA: each row is a Smith-Waterman alignment to a repeat sequence. The columns provide various characteristics of each repeat match including their strand orientation (+/C), repeat class families, repeat positions and other sequence statistics.

File 4 (.masked)

This file returns the query with repeat and/or low complexity sequence nucleotides replaced by an 'X' character. 'X' is used rather than 'N' to distinguish masked repeats from ambiguous bases (N).

File 5 (.cat)

This file details conserved regions and potential repetitive element alignments found between the query and the fungal repat genome library, including corresponding aligned sequences (scored with symbols) and various summary statistics. For each alignment, it provides the repeat status, REPAT clustering method (Grouper, Piler or Recon) and repeat class.

What repeats are found?

Repeats found belonged to four well-defined classes [1]. Out of the four classes, 3 LINEs and 2 LTRs were identified in the 7.97% of the query sequence. There were 21 unclassified repeat classes which made up 35.13% of the sequence.

Task 2: Homology Methods

BLASTn: Determining the FASTA's evolutionary relationships using the non-redundant nucleotide database (nr) [2].

All E-value significant BLASTn alignments belonged to the *Blumeria Graminis* (*B.graminis)* species, evident from their distance tree (Figure 1A). There are 8 *forma speciales* (f.sp*)* of *B.graminis* (including *tritici and hordei*) [3]. Close matches were predominantly to *tritici* (Figure 1A), excluding one (closest) identity match to *hordei* (Figure 1B). This could be repeat-biased.

BLASTn (re-run): Using the repeat-masked FASTA to avoid spurious or repeat-biased matches [4].

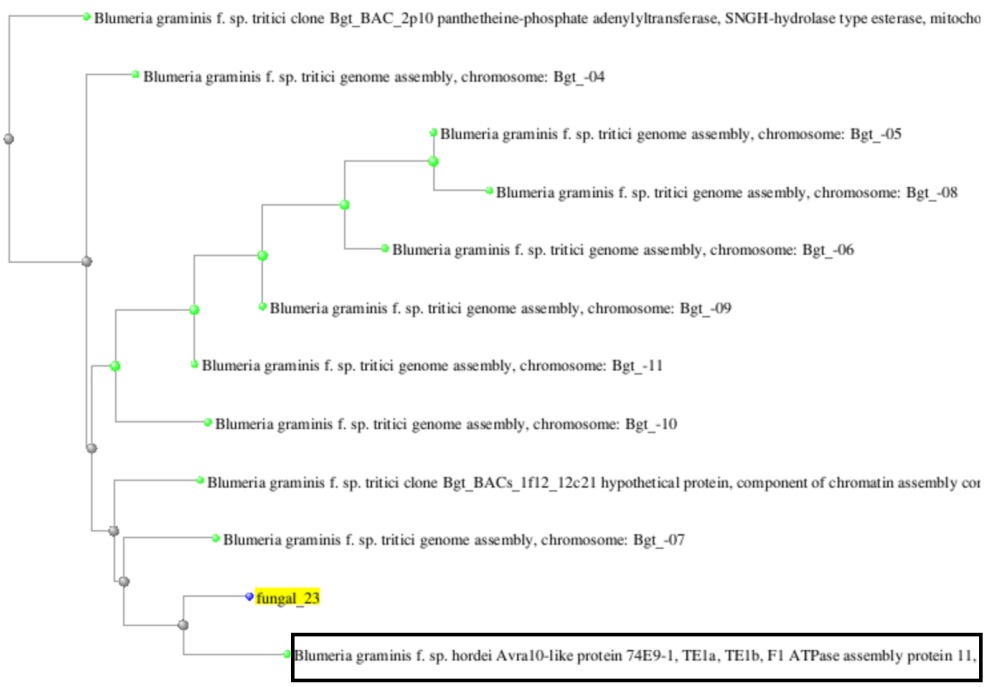
After the first ~400 nucleotides, all alignments were from *B.graminis tritici* f.sp, evident from taxonomy (Figure 2A). The *tritici* alignment had a high query cover, spanning 90% of the sequence, with a 92% identity score (Figure 2B). This confirms the proposed possibility of a repeat-biased *hordei* alignment.

BLASTx: Comparing the translated FASTA to the non-redundant protein sequence database within the *Blumeria* Genus.

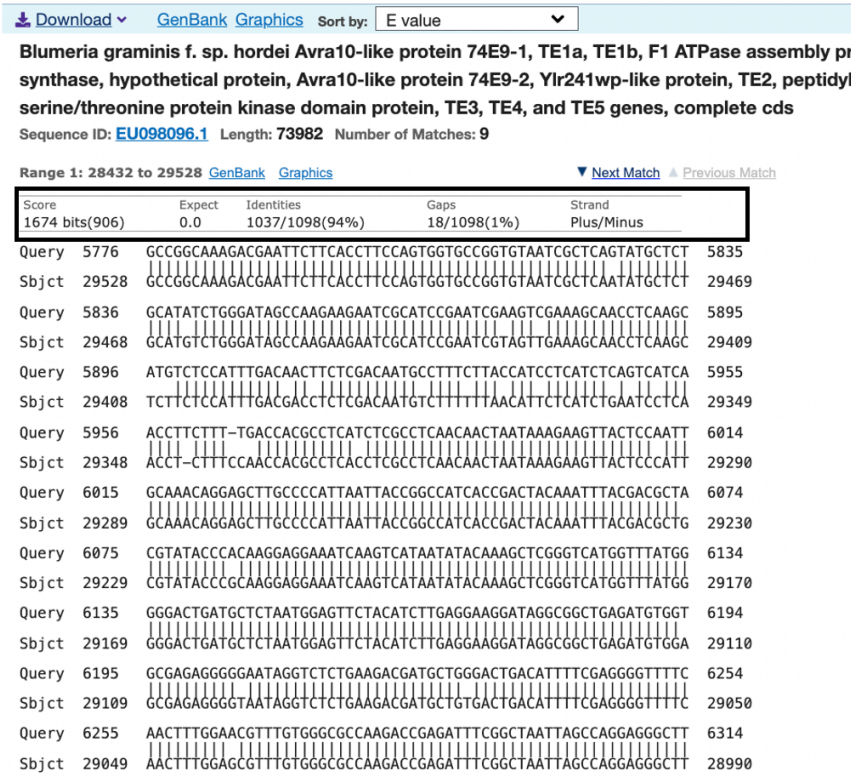
The *B.graminis hordei* f.spp results had a higher cumulative score than *tritici* but this was likely repeat-biased as there were far less *hordei* alignments (Figure 2A). Most high confidence proteins had overlapping sequences from both *hordei* and *tritici* f.sp but, whilst many *hordei* proteins had functional annotations, the equivalent proteins in *tritici* were mostly hypothetical (Figure 3B).

Figure 1: BLASTn results of the fungal_23 FASTA sequence against the entire nucleotide database. A. Distance Tree (adapted). There are many sequences which produce sequence alignments from the *tritici* f.sp of *Blumeria Graminis* but the sequences closest in relationship to the FASTA query belongs to the *hordei* f.sp (enclosed in a black box). **B. The top identity BLASTn match from the *Blumeria Graminis f.sp. hordei* (adapted)**. This sequence corresponds to the closely related alignment in the previous image. Identity and alignment scores are very high and enclosed by the black box.

A.



B.



Download GenBank Graphics Sort by: E value

Blumeria graminis f. sp. hordei Avra10-like protein 74E9-1, TE1a, TE1b, F1 ATPase assembly pr synthase, hypothetical protein, Avra10-like protein 74E9-2, Ylr241wp-like protein, TE2, peptidyl serine/threonine protein kinase domain protein, TE3, TE4, and TE5 genes, complete cds

Sequence ID: EU098096.1 Length: 73982 Number of Matches: 9

Range 1: 28432 to 29528 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
1674 bits(906)	0.0	1037/1098(94%)	18/1098(1%)	Plus/Minus

```
Query  5776   GCCGGCAAAGACGAATTCTTCACCTTCCAGTGGTGCCGGTGTAATCGCTCAGTATGCTCT  5835
              ||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||
Sbjct  29528  GCCGGCAAAGACGAATTCTTCACCTTCCAGTGGTGCCGGTGTAATCGCTCAATATGCTCT  29469

Query  5836   GCATATCTGGGATAGCCAAGAAGAATCGCATCCGAATCGAAGTCGAAAGCAACCTCAAGC  5895
              |||| ||||||||||||||||||||||||||||||||| ||| |||||||||||||||||
Sbjct  29468  GCATGTCTGGGATAGCCAAGAAGAATCGCATCCGAATCGTAGTTGAAAGCAACCTCAAGC  29409

Query  5896   ATGTCTCCATTTGACAACTTCTCGACAATGCCTTTCTTACCATCCTCATCTCAGTCATCA  5955
                  ||||||||||||| || ||||||||||| |||| ||| ||| ||||||| | || |||
Sbjct  29408  TCTTCTCCATTTGACGACCTCTCGACAATGTCTTTTTAACATTCTCATCTGAATCCTCA  29349

Query  5956   ACCTTCTTT-TGACCACGCCTCATCTCGCCTCAACAACTAATAAAGAAGTTACTCCAATT  6014
              |||| ||||    ||||||||||| ||||||||||||||||||||||||||||||| |||
Sbjct  29348  ACCT-CTTTCCAACCACGCCTCACCTCGCCTCAACAACTAATAAAGAAGTTACTCCCATT  29290

Query  6015   GCAAACAGGAGCTTGCCCCATTAATTACCGGCCATCACCGACTACAAATTTACGACGCTA  6074
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  29289  GCAAACAGGAGCTTGCCCCATTAATTACCGGCCATCACCGACTACAAATTTACGACGCTG  29230

Query  6075   CGTATACCCACAAGGAGGAAATCAAGTCATAATATACAAAGCTCGGGTCATGGTTTATGG  6134
              |||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  29229  CGTATACCCGCAAGGAGGAAATCAAGTCATAATATACAAAGCTCGGGTCATGGTTTATGG  29170

Query  6135   GGGACTGATGCTCTAATGGAGTTCTACATCTTGAGGAAGGATAGGCGGCTGAGATGTGGT  6194
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  29169  GGGACTGATGCTCTAATGGAGTTCTACATCTTGAGGAAGGATAGGCGGCTGAGATGTGGA  29110

Query  6195   GCGAGAGGGGGAATAGGTCTCTGAAGACGATGCTGGGACTGACATTTTCGAGGGGTTTTC  6254
              |||||||||| |||||||||||||||||||||||||| |||||||||||||||||||||||
Sbjct  29109  GCGAGAGGGGTAATAGGTCTCTGAAGACGATGCTGTGACTGACATTTTCGAGGGGTTTTC  29050

Query  6255   AACTTTGGAACGTTTGTGGGCGCCAAGACCGAGATTTCGGCTAATTAGCCAGGAGGGCTT  6314
              ||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  29049  AACTTTGGAGCGTTTGTGGGCGCCAAGACCGAGATTTCGGCTAATTAGCCAGGAGGGCTT  28990
```

Figure 2: BLASTn of repeat-masked fungal_23 FASTA sequence against the entire nucleotide database. A. Taxonomy results from the output. BLAST has identified the sequence to be in the *Blumeria Graminis* species. However, unlike the search of the unmasked FASTA, there are no results for the *hordei* f.sp whereas there is one *tritici* alignment with very high score (enclosed in a black rectangle). Notably the rest of the hits are only aligned to the region at the start of the sequence which is highly conserved across related ascomycete fungi. The *Botrytis* species has been enclosed in a square box because it is a closely related species relevant to the next section. **B. Graphic depicting the high score *tritici* genome assembly alignment to the masked query sequence.** The length of the sequence has been reduced to ~7500 nucleotides (from ~13,800) since BLASTn deletes the X characters out of the masked query. There is a highly significant nucleotide match to the *tritici* f.sp genome assembly of the *Blumeria Graminis* species that spans 90% of the sequence with 92% identity.

A.

Organism	Blast Name	Score	Number of Hits
Eukaryota	eukaryotes		101
. Opisthokonta	eukaryotes		80
. . Dikarya	fungi		78
. . . Ascomycota	ascomycete fungi		70
. . . . saccharomyceta	ascomycete fungi		67
. . . . . leotiomyceta	ascomycete fungi		64
. . . . . sordariomyceta	ascomycete fungi		38
. . . . . Leotiomycetes	ascomycete fungi		9
. . . . . Blumeria graminis f. sp. tritici	powdery mildews	4543	1
. . . . . Mollisia scopiformis	ascomycete fungi	363	1
. . . . . Hyaloscypha bicolor E	ascomycete fungi	348	1
. . . . . Allantophomopsis cytisporea	ascomycete fungi	331	1
. . . . . Allantophomopsis lycopodina	ascomycete fungi	327	1
. . . . . Amorphotheca resinae ATCC 22711	ascomycete fungi	326	1
. . . . . Drepanopeziza brunnea f. sp. 'multigermtubi' MB_m1	ascomycete fungi	195	1
. . . . . Botrytis byssoidea	ascomycete fungi	182	1
. . . . . Botrytis deweyae	ascomycete fungi	178	1

B.

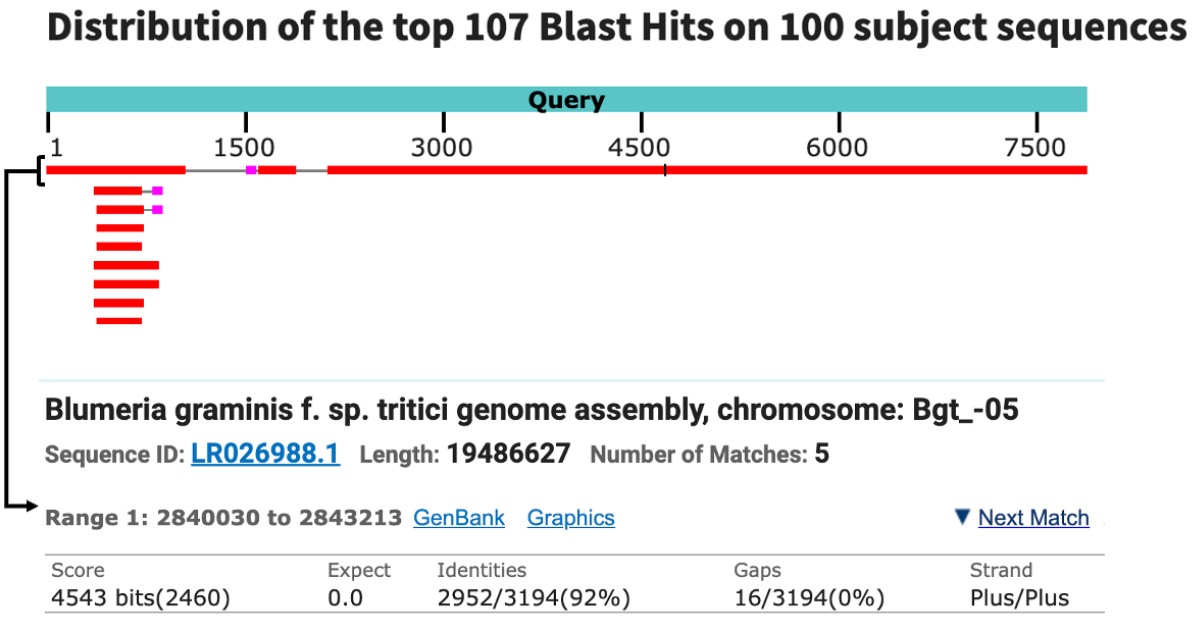


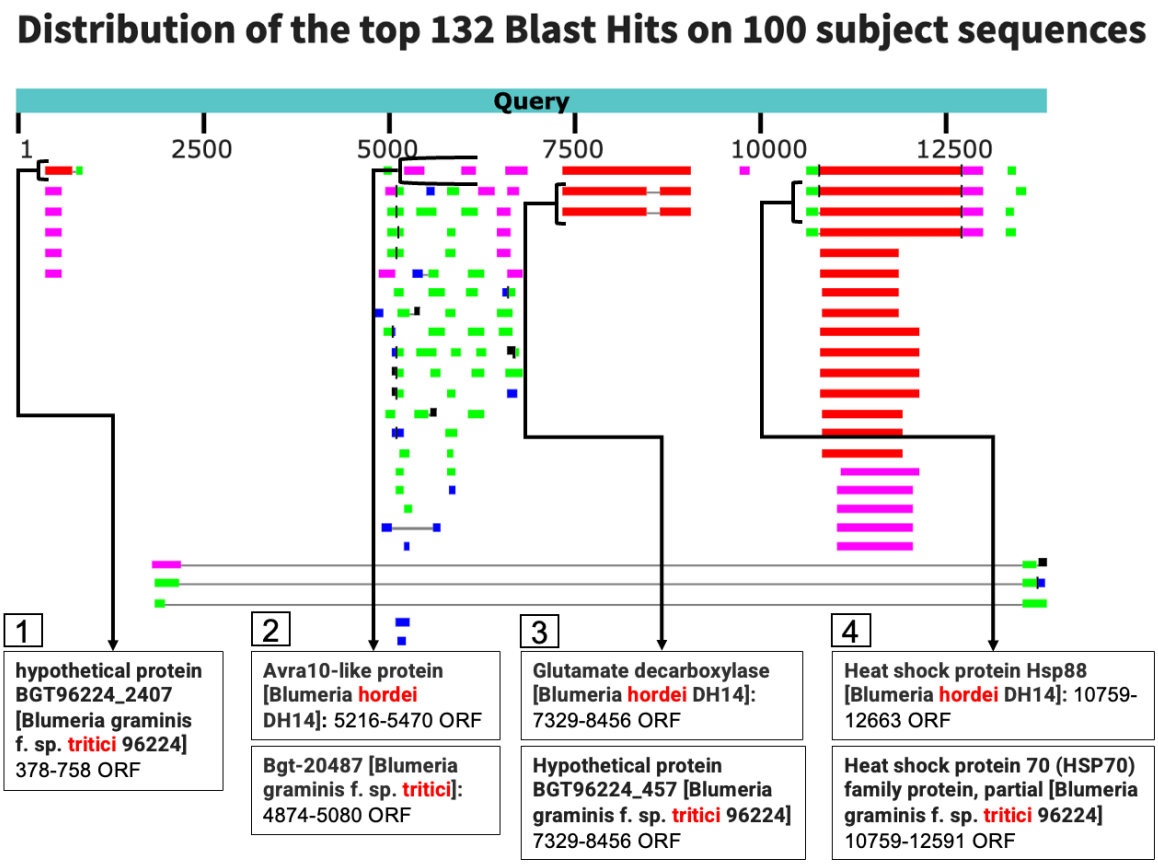
Figure 3: BLASTx results. A. This image shows the general taxonomy of all the hits (adapted). The search was restricted to the *Blumeria* Genus. Since this genus is monospecific, all hits are from *Blumeria Graminis.* Scores corresponding to various f.sp of this species can be seen. All scores are very high but, although most hits are for the *tritici* f.sp, the highest scores are achieved by alignment to the *hordei* f.sp (highlighted in yellow). **B. This image shows the distribution of the top 132 BLAST hits on 100 subject sequences (adapted).** Four key protein alignments were identified and annotated of which 3 had high confidence alignment scores in red (moderate confidence scores are pink). Where annotations were available, these were added below the alignment graphic and connected to the corresponding alignment(s) using a black arrow. The open reading frames (ORFs) were also included for each alignment.

A.

100 sequences selected

Organism	Blast Name	Score	Number of Hits	Description
Blumeria	powdery mildews		117	
. Blumeria hordei	powdery mildews	1184	9	Blumeria hordei hits
. Blumeria hordei DH14	powdery mildews	1114	12	Blumeria hordei DH14 hits
. Blumeria graminis f. sp. tritici 96224	powdery mildews	1088	8	Blumeria graminis f. sp. tritici 96224 hits
. Blumeria graminis f. sp. tritici	powdery mildews	1088	74	Blumeria graminis f. sp. tritici hits
. Blumeria graminis f. sp. triticale	powdery mildews	1087	14	Blumeria graminis f. sp. triticale hits

B.



Task 3: *Ab initio* Methods & Task 4: *Ab initio* Gene Prediction – fgenesh

AUGUSTUS

Augustus's *ab initio* model has ~100 species-specific training sets but for closely related species 'usually only one version is necessary' [5]. Using the fungal phylogenetic tree (Figure 4A), the *Botrytis Cineria* species is the taxonomically closest available training set to *B.graminis*. The algorithm identified 7 total transcripts from 32 exons and 4 unique genes (Figure 4B) [3].

FGENESH

FGENESH uses pre-trained model gene-finding parameters (506 in total) [6]. Model parameters were available for *B.Graminis* which predicted 13 protein-coding exons (Figure 5A, highlighted), corresponding to 5 predicted gene transcripts (Figure 5B).

Comparison

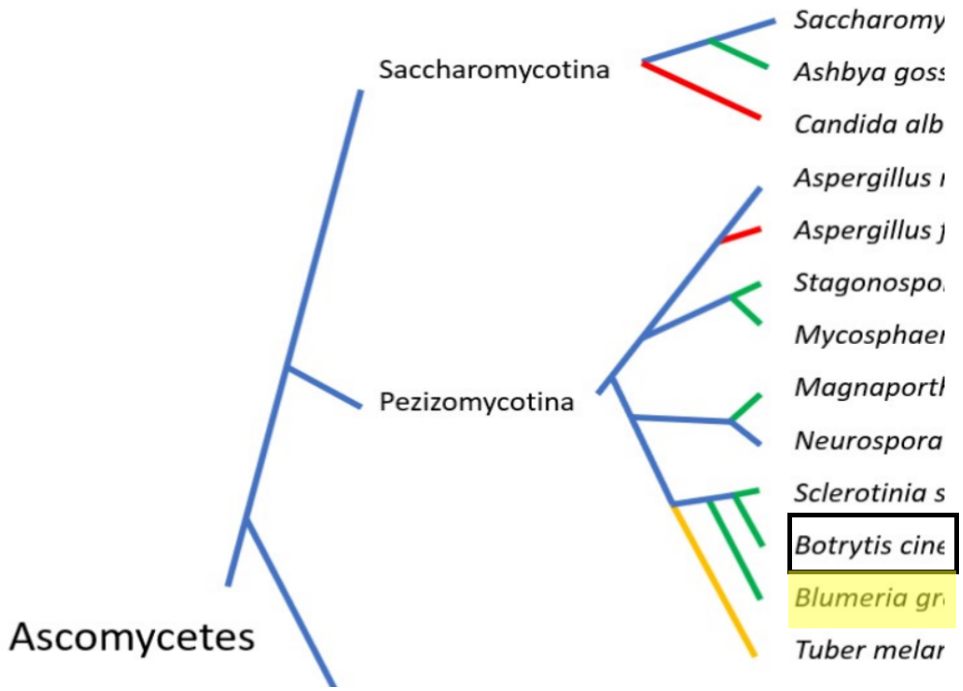
The main differences were between predicted gene open reading frames (ORFs) (Table 1) and transcript-to-gene counts (Augustus found 4 genes with 7 transcripts whereas FgenesH found 5 genes with 5 transcripts).

Augustus generating more transcripts for less genes suggests greater predictive uncertainty. This arises due to its lack of *B.Graminis* parameters for its generalised Hidden Markov Model (HMM), which was re-trained for the query sequence using *Botrytis Cineria* parameters.

FgenesH also uses a HMM but with a much larger database of pre-trained models (506) including the *B.graminis* species, which provides a better reference genome and facilitates a more refined feature output for the fungal_23 query FASTA.

Figure 4: AUGUSTUS results. A. Fungal phylogenetic tree used to guide the training set chosen for the AUGUSTUS search. *Blumeria Graminis* (highlighted*)* is not an available training set on AUGUSTUS. Using the phylogenetic tree, the *Botrytis Cineria* species is the most closely related species available on Augustus (enclosed in a black box). **B. The graphical browsable results from Augustus (adapted).** Augustus has identified 4 different protein coding sequences with 7 transcripts from 32 exons. Three of the four genes have two transcripts (t1 and t2). Each gene has been annotated with exon counts per transcript and their open reading frame (ORF).

A.



B.

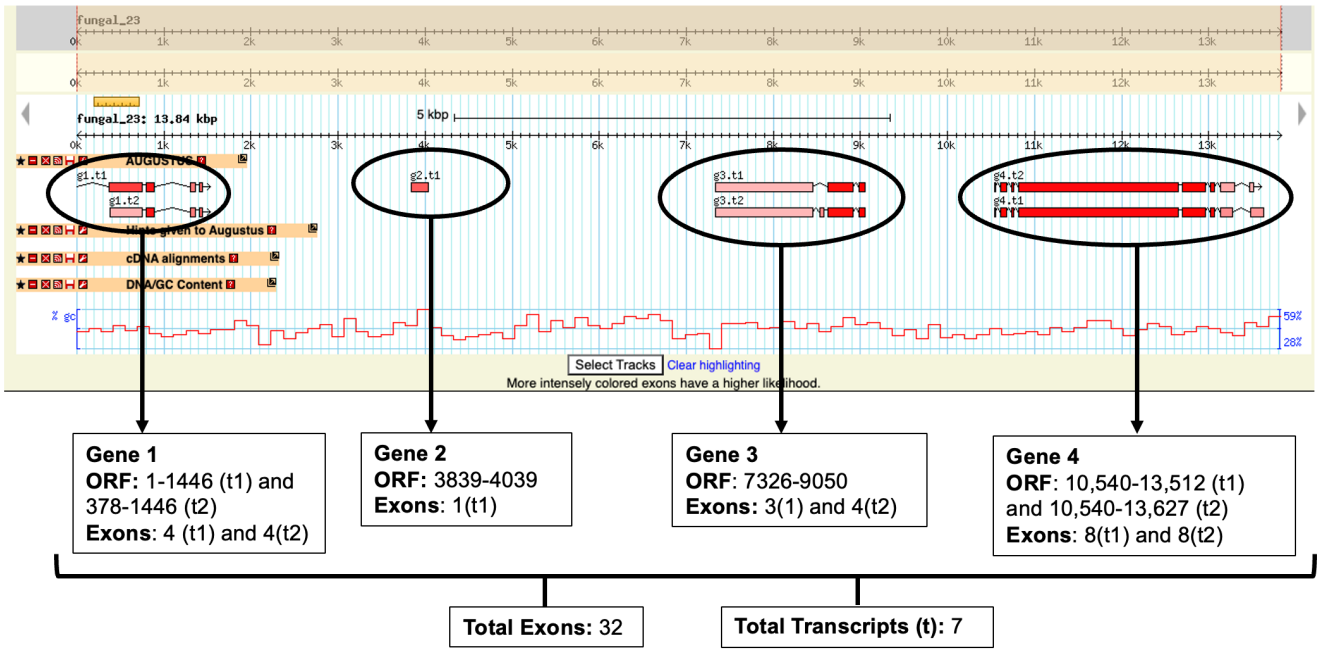
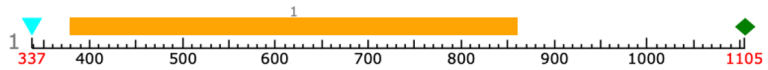
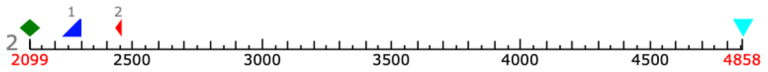
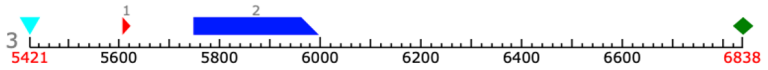
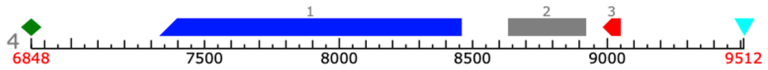
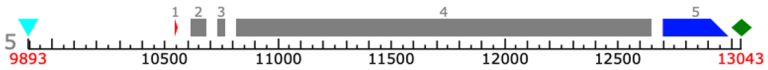


Figure 5: Graphical output results of FgenesH (adapted): 'CD' annotations describe exon predictions from FgenesH: the **CDSf** is the first exon of the gene sequence (with the start codon), **CDSi** is an internal exon and **CDSl** is the last coding exon (with the stop codon). **CDSo** is used for genes with only one coding region (CDS with start and stop codon). The coding sequences of 5 predicted genes have been organised into a table with their corresponding FgenesH graphics. Exon counts for each gene identified are included in one column, totalling 13 protein-coding exons. The coding strand is also provided, where + refers to the forward strand and – the reverse strand.

Key	CDSf CDSi CDSl CDSo PolA TSS	Exons: **(13 total)**	Strand **(+/-)**
Gene 1: CDSo (378-860)		**1**	**+**
Gene 2: CDSl (2223-2429), CDSf (2431-2454)		**2**	**-**
Gene 3: CDSL 5747-5995, CDSf (5605-5619)		**2**	**+**
Gene 4: CDSL (7326- 8456), CDSi (8628- 8918)		**3**	**-**
Gene 5: CDSL 12695- 12985, CDSi 10613-12645, CDSf (10540-10545)		**5**	**+**

Task 5: Protein-based annotations

Ab initio gene predictors generate conceptual translations of putative proteins within genomic sequences without experimental evidence (Figures 4 and 5) [5][6]. BLASTx can suggest experimentally characterised proteins and predicted protein annotations inferred from known and hypothetical proteins within its non-redundant protein sequence database (Figure 3) [2].

Table 1 compares open reading frames (ORFs) between *ab initio* and BLASTx predicted proteins encoded by the fungal_23 FASTA. BLAST annotations were suggested for *ab initio* ORFs using BLASTp (see Appendix); this facilitated grouping of *ab initio* and BLASTx sequences. Similar ORFs were highlighted.

Putative Gene Model

FgenesH predicted 13 protein coding exons corresponding to 5 proteins which were mapped onto 4 BLASTx protein annotations (Table 1). Multiple overlapping ORFs between FgenesH predictions and BLASTx alignments supported the previous conclusion that FgenesH is the more accurate *ab initio* prediction program for the query FASTA sequence.

Protein 1 was hypothetical but had a strong consensus ORF between BLASTx and FgenesH, unlike protein(s) 2 which was excluded from further study. Proteins 3 and 4 both had robust consensus ORFs across software; however, whilst Protein 4 was only annotated for the *B.graminis hordei* f.sp, Protein 3 was annotated for both *hordei* and *tritici* (Table 1).

Using BLASTx annotations and FgenesH genomic structure predictions, a putative gene model for *B.graminis* was hypothesised (Table 2). Given the repeat-masked BLASTn results (Figure 2B), this model used feature annotations from the *B.graminis f.sp.tritici* genome. All three features incorporated had highly significant E-Values (~0) and >90% sequence identity to the query FASTA (Figure 6).

The three features incorporated in the model had been sequenced and uploaded to the BLAST database by a recent paper that attempted to draft the *B.graminis f.sp.tritici* genome [7]. Although Protein 3 could be inferred as glutamate decarboxylase from the *hordei DH14* annotation with an equivalent ORF (Table 1), Protein 1 and 3 were both functionally hypothetical for this genome. On the other hand, Protein 4 had a robust functional annotation as a Heat Shock Protein 70 (Hsp70) family member. Hsp70 proteins are ubiquitous, highly conserved ATP-dependent protein folding chaperones (~70-kDa) which are essential for life in virtually all living organisms [8].

BLAST annotation(s) – see Figure 3B	**BLASTx**	**FgenesH**	**Augustus**
Protein 1: *hypothetical protein BGT96224_2407 [Blumeria graminis f. sp. tritici 96224*	378-758	378-860	378-1446
Protein(s) 2: *Avra10-like protein [Blumeria hordei DH14]* **and** *Bgt-20487 [Blumeria graminis f. sp. tritici]*	5216-5470 **and** 4874-5080	2223-2429 5747-5995	3839-4039
Protein 3: *Glutamate decarboxylase [Blumeria hordei DH14]* **or** *Hypothetical protein BGT96224_457 [Blumeria graminis f. sp. tritici 96224]*	7329-8456	7326-8456	7326-9050
Protein 4: *Heat shock protein Hsp88 [Blumeria hordei DH14]* **or** *Heat shock protein 70 (HSP70) family protein, partial [Blumeria graminis f. sp. tritici 96224]*	10759-12663 **or** 10759-12591	10613-12985	10,540-13,627

Table 1: Table to compare open reading frames of proteins predicted by *ab initio* software or retrieved by BLASTx from its non-redundant database. Since the software produced a different number of proteins, this table groups the proteins into four groups using BLASTx annotations. These annotations can be viewed in Figure 3B where a graphic of the BLASTx output was manually annotated: proteins 1,3 and 4 were all high score alignments whereas Protein 2 is a group of proteins with ambiguous alignments and annotations. Significant E-value BLASTx alignments from both *hordei* and *tritici* f.sp of *B.graminis* are provided as annotations, separated by 'or' when they both correspond to similar alignments. For each protein annotation, similar open reading frames between BLASTx, FgenesH and Augustus are highlighted. They are highlighted in yellow for Protein 1, pink for Protein 2 and green for Protein 3. Proteins with consensus, highlighted ORFs were used to build a putative gene model in the next step (Table 2).

Protein (BLASTx)	Coding region (BLASTx)	Exons (FgenesH)	Strand	Inside repeat?
Protein 1: *Hypothetical protein BGT96224_2407 [Blumeria graminis f. sp. tritici 96224*	378-758	1	Forward	No
Protein 3: *Hypothetical protein BGT96224_457 [Blumeria graminis f. sp. tritici 96224]*	7329-8456	3	Reverse	No
Protein 4: *Heat shock protein 70 (HSP70) family protein, partial [Blumeria graminis f. sp. tritici 96224]*	10759-12591	5	Forward	No

Table 2: BLASTx Putative Gene Model for FASTA query sequence (*Blumeria graminis f. sp. tritici* species). Proteins 1, 3 and 4 from Table 1 were used to build a putative gene model for the FASTA query sequence due to their high-scoring BLASTx alignments (see Figure 3B). Protein annotations corresponding to the *tritici* f.sp of *B.graminis* were retrieved from BLASTx results in table 1 (verified for sequences with BLASTp in the appendix). FgenesH genomic structure information was retrieved from figure 5B for each feature. This information included open reading frames (ORFs), exon counts and the coding strand orientation. Each open reading frame was checked against RepeatMaster results to confirm that it was not in a repeat sequence.

Protein Annotations	Organism				E-Value	Identity (%)		
unnamed protein product [Blumeria hordei]	Blumeria hordei	1184	1368	17%	0.0	95.95%	744	SZF02134.1
heat shock protein Hsp88 [Blumeria hordei DH14]	Blumeria hordei DH14	1114	1297	17%	0.0	87.93%	711	CCU74275.1
heat shock protein 70 (HSP70) family protein [Blumeria graminis f. sp. tritici 96224]	Blumeria graminis f. sp. tritici ...	1088	1268	16%	0.0	90.72%	709	EPQ65459.1
BgTH12-01506 [Blumeria graminis f. sp. triticale]	Blumeria graminis f. sp. triticale	1087	1272	16%	0.0	90.87%	711	CAD6501254.1
unnamed protein product [Blumeria hordei]	Blumeria hordei	979	1060	12%	0.0	85.15%	530	SZF02046.1
glutamate decarboxylase [Blumeria hordei DH14]	Blumeria hordei DH14	786	1021	11%	0.0	100.00%	511	CCU74163.1
hypothetical protein BGT96224_457 [Blumeria graminis f. sp. tritici 96224]	Blumeria graminis f. sp. tritici ...	753	974	11%	0.0	94.41%	511	EPQ63547.1
ATP binding protein [Blumeria graminis f. sp. tritici 96224]	Blumeria graminis f. sp. tritici ...	244	244	7%	5e-68	38.76%	648	EPQ67700.1
heat shock protein 70 [Blumeria graminis f. sp. tritici]	Blumeria graminis f. sp. tritici	244	244	7%	5e-68	38.76%	648	AFJ79448.1
hypothetical protein BGT96224_2407 [Blumeria graminis f. sp. tritici 96224]	Blumeria graminis f. sp. tritici ...	201	269	3%	1e-64	98.43%	155	EPQ66886.1

Figure 6: BLASTx scores and annotations of Table 1 proteins selected to build the putative gene model. Highlighting (colour) corresponds to the same highlighted protein annotation in Table 1 and Table 2. The output has been labelled for protein annotations, organism, E-value and identity. The E-value scores for highlighted annotations are very low and identity scores are very high, validating the strength of these alignments to the FASTA query sequence.

References

[1] Smit, A F, and A D Riggs. “Tiggers and DNA transposon fossils in the human genome.” *Proceedings of the National Academy of Sciences of the United States of America* vol. 93,4 (1996): 1443-8. doi:10.1073/pnas.93.4.1443

[2] McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W20-5. doi: 10.1093/nar/gkh435. PMID: 15215342; PMCID: PMC441573.

[3] Troch V, Audenaert K, Wyand RA, Haesaert G, Höfte M, Brown JK. Formae speciales of cereal powdery mildew: close or distant relatives? Mol Plant Pathol. 2014 Apr;15(3):304-14. doi: 10.1111/mpp.12093. Epub 2014 Jan 6. PMID: 24286122; PMCID: PMC6638862.

[4] A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org

[5] 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

[6] Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. 2006;7 Suppl 1(Suppl 1):S10.1-12. doi: 10.1186/gb-2006-7-s1-s10. Epub 2006 Aug 7. PMID: 16925832; PMCID: PMC1810547.

[7] Wicker, T., Oberhaensli, S., Parlange, F. *et al.* The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. *Nat Genet* **45**, 1092–1096 (2013). https://doi.org/10.1038/ng.2704

[8] Sharma D, Masison DC. Hsp70 structure, function, regulation and influence on yeast prions. Protein Pept Lett. 2009;16(6):571-81. doi: 10.2174/092986609788490230. PMID: 19519514; PMCID: PMC2746719.

Appendix

Augustus Genes:

Gene g1: **hypothetical protein BGT96224_2407 [Blumeria graminis f. sp. tritici 96224]**

- t1:AVMQIFVKTLTGKTITLEVESSDTIDNVKTKIQDKEGIPPDQQRLIFAGKQLEDGRTLADYNIQKESTLHLVLRLRGGAKKRKKKVYTTPKKIKHKRKKTKLAVLKYYKVDGDGKIERLRRECPSTDCGAGVFMAAMQDRQYCGRCHLTYIFDDAANLKIDHVNQMDYVRMFEETTCSVSPELIDYL
- t2:MQIFVKTLTGKTITLEVESSDTIDNVKTKIQDKEGIPPDQQRLIFAGKQLEDGRTLADYNIQKESTLHLVLRLRGGAKKRKKKVYTTPKKIKHKRKKTKLAVLKYYKVDGDGKIERLRRECPSTDCGAGVFMAAMQDRQYCGRCHLTYIFDDAANLKIDHVNQMDYVRMFEETTCSVSPELIDYL

Gene g2: **Unknown sequence**

- t1:MWSRSQNITPGEVAIRWCQSYAGSLGQKAAGVLARAVYMASAPHSPLFIARARREVKAKYEASVAL

Gene g3: **Glutamate decarboxylase (Blumeria hordei DH14)**

- t1:MESNHYASGEPQGQPDHASELEDLLHAVQALIIPFVRSADESVACKSMTGSVPVAPGASVRTSLVESLSPDVLRETMNFNLPLTGQGKVGLLRAIEQVLHYSVNTGDQGFMDKLYAGTSPVHVFKVSPALSVIEKLTAKALAVQFGFDGPDSGGISTQSGSASNATSIVIARNCLYPSTKKEGNASLKLVIFTSTDGHYSLEKAAQVCGLGSDNVCLVPSDSDGRMIPEELDRLVQIVKAEGRTPFYVNATAGTTVLGSYDPVEAISLVCKKHHLWLHVDASWGGPVIFSRRLKYKMSGSHLADSLAINPHKMMGVPVTCSFLLTPDLRRFHSANSIAADYLFHTESTRSEVWDLAELTIQCGRRADSLKLALSWIYYGTEGFEQQIEYAFSVAAYMAAQIESNPHFHLVSENPPPCLQVCFYYTGSAGPSTSKNLNEQMTSCLAERLVPRGFMVDYAPGKWGSFLRVVVNRETRIATVDSLLKMIEEIGSTIVIK
- t2:MESNHYASGEPQGQPDHASELEDLLHAVQALIIPFVRSADESVACKSMTGSVPVAPGASVRTSLVESLSPDVLRETMNFNLPLTGQGKVGLLRAIEQVLHYSVNTGDQGFMDKLYAGTSPVGVVSELVLAILNTNVHVFKVSPALSVIEKLTAKALAVQFGFDGPDSGGISTQSGSASNATSIVIARNCLYPSTKKEGNASLKLVIFTSTDGHYSLEKAAQVCGLGSDNVCLVPSDSDGRMIPEELDRLVQIVKAEGRTPFYVNATAGTTVLGSYDPVEAISLVCKKHHLWLHVDASWGGPVIFSRRLKYKMSGSHLADSLAINPHKMMGVPVTCSFLLTPDLRRFHSANSIAADYLFHTESTRSEVWDLAELTIQCGRRADSLKLALSWIYYGTEGFEQQIEYAFSVAAYMAAQIESNPHFHLVSENPPPCLQVCFYYTGSAGPSTSKNLNEQMTSCLAERLVPRGFMVDYAPGKWGSFLRVVVNRETRIATVDSLLKMIEEIGSTIVIK

Gene g4: **heat shock protein Hsp88 [Blumeria hordei DH14]**

- t1:MSVIGVDFGTAGTVIAVARNRGVDVITNEVSNRSTPSLVGFGPKSRYLGESAKTQEISNL KNTVASLKRLAGRCLNDPDIQIEQNFISAPLVDINGQVGAEVTYLNRKERFTSTQLISMFLSK IKATASAELKLPVSDLVMSVPAWFTDAQRRSLMDAAEIAGLKLLRLMNDTTAAALGYGITKL DLPTAEENPRRVAFIDIGHSNYTCSIVEFRKGELTVKSTAYDRHFGGRDFDRALVEHLATEF KESYKIDIKSNAKAMVRVNAAAEKLKKILSANQQAPLNIESLMDDIDVATMVTREGLEKLVE PLLNRVHVPLEQALADAKLKVEDIDVIELVGGCTRVPALKERIQKFFNKTLSFTLNQDEAIAR GCAFSCAILSPVFRVRDFSIHDIVNYPIEFTWEKSPDIPDEDTSLTVFNQGNLMPSTKILTFY RKEPFNLEAKYATPEKLPGKTKPWIGNFSVKGVKADAKDDFMICKLKARLNLHGILNVDQG YYVEEVEVEEPIPEDPEKEPEVSFELNPFHVRSSSEERAVKRVRTGQPVLTSTKAMDIEDK PKTRKVKSQVRKGDLPISSATSGLDDSSKSASAEQEAAMIMEDKLVSDTEEKNELEAYIYE MRNKLDDQYSEFASEEEKLKLNERLTTTEDWLYDEGEDATKATYIAKMDEIRMLAGPITQR HFDKVEEERAAAQAISDAEAAKKRELAEAAKAAALAQEAAMSKDEVMAEADSSKNSEGLE PELLKRTKNGGYAGDNEKQWSNFKLLGRYPVYVMILRTQKKYDLIVHSDDTIYQSYSSMS VTEPYSSLTLDHGGESILCIPEAYPAPHTEMAAGIKSASDVTTKGDID
- t2:MSVIGVDFGTAGTVIAVARNRGVDVITNEVSNRSTPSLVGFGPKSRYLGESAKTQEISNL KNTVASLKRLAGRCLNDPDIQIEQNFISAPLVDINGQVGAEVTYLNRKERFTSTQLISMFLSK IKATASAELKLPVSDLVMSVPAWFTDAQRRSLMDAAEIAGLKLLRLMNDTTAAALGYGITKL DLPTAEENPRRVAFIDIGHSNYTCSIVEFRKGELTVKSTAYDRHFGGRDFDRALVEHLATEF KESYKIDIKSNAKAMVRVNAAAEKLKKILSANQQAPLNIESLMDDIDVATMVTREGLEKLVE PLLNRVHVPLEQALADAKLKVEDIDVIELVGGCTRVPALKERIQKFFNKTLSFTLNQDEAIAR GCAFSCAILSPVFRVRDFSIHDIVNYPIEFTWEKSPDIPDEDTSLTVFNQGNLMPSTKILTFY RKEPFNLEAKYATPEKLPGKTKPWIGNFSVKGVKADAKDDFMICKLKARLNLHGILNVDQG YYVEEVEVEEPIPEDPEKEPEVSFELNPFHVRSSSEERAVKRVRTGQPVLTSTKAMDIEDK PKTRKVKSQVRKGDLPISSATSGLDDSSKSASAEQEAAMIMEDKLVSDTEEKNELEAYIYE MRNKLDDQYSEFASEEEKLKLNERLTTTEDWLYDEGEDATKATYIAKMDEIRMLAGPITQR HFDKVEEERAAAQAISDAEAAKKRELAEAAKAAALAQEAAMSKDEVMAEADSSKNSEGLE PELLKRTKNGGYAGDNEKQWSNFKLLGRYPVYVMILRTQKKYDLIVHSDDTIYQSYSSVSD RRYTHDLSDVSDRALLITNS

<u>FgenesH Genes:</u>

Gene 1: **hypothetical protein BGT96224_2407 [Blumeria graminis f. sp. tritici 96224]**

- MQIFVKTLTGKTITLEVESSDTIDNVKTKIQDKEGIPPDQQRLIFAGKQLEDGRTLADYNIQK ESTLHLVLRLRGGAKKRKKKVYTTPKKIKHKRKKTKLAVLKYYKVDGDGKIERLRRECPST DVSPLSVVSHILILISAVPVFSWLLCRTVNIVDDAI

Gene 2: **N/a**

- MSFAHVAFRLYIVPPESLIVEHIGLLSNPDHVF

Gene 3: **Sucrose transporter (Blumeria hordei DH14)**

- MGQLVRKSRTGYLLGLPAKTNSSPSSGAGVIAQYALHIWDSQEESHPNRSRKQPQACLHL TTSRQCLSYHPHLSHQPSFDHASSRLNN

Gene 4: **hypothetical protein BGT96224_457 [Blumeria graminis f. sp. tritici 96224]**

- MESNHYASGEPQGQPDHASELEDLLHAVQALIIPFVRSADESVACKSMTGSVPVAPGASV RTSLVESLSPDVLRETMNFNLPLTGQGKVGLLRAIEQVLHYSVNTGDQGFMDKLYAGTSP VHVFKVSPALSVIEKLTAKALAVQFGFDGPDSGGISTQSGSASNATSIVIARNCLYPSTKKE GNASLKLVIFTSTDGHYSLEKAAQVCGLGSDNVCLVPSDSDGRMIPEELDRLVQIVKAEGR TPFYVNATAGTTVLGSYDPVEAISLVCKKHHLWLHVDASWGGPVIFSRRLKYKMSGSHLA DSLAINPHKMMGVPVTCSFLLTPDLRRFHSANSIAADYLFHTESTRSEVWDLAELTIQCGR RADSLKLALSWIYYGTEGFEQQIEYAFSVAAYMAAQIESNPHFHLVSENPPPCLQVCFYYT GSAGPSTSKNLNEQMTSCLAERLVPRGFMVDYAPGKWGSFLRVVVNRETRIATVDSLLK MIEEIGSTIVIK

Gene 5: **heat shock protein 70 (HSP70) family protein, partial [Blumeria graminis f. sp. tritici 96224]**

- MSVIGVDFGTAGTVIAVARNRGVDVITNEVSNRSTPSLVGFGPKSRYLGESAKTQEISNLK NTVASLKRLAGRCLNDPDIQIEQNFISAPLVDINGQVGAEVTYLNRKERFTSTQLISMFLSKI KATASAELKLPVSDLVMSVPAWFTDAQRRSLMDAAEIAGLKLLRLMNDTTAAALGYGITKL DLPTAEENPRRVAFIDIGHSNYTCSIVEFRKGELTVKSTAYDRHFGGRDFDRALVEHLATEF KESYKIDIKSNAKAMVRVNAAAEKLKKILSANQQAPLNIESLMDDIDVATMVTREGLEKLVE PLLNRVHVPLEQALADAKLKVEDIDVIELVGGCTRVPALKERIQKFFNKTLSFTLNQDEAIAR GCAFSCAILSPVFRVRDFSIHDIVNYPIEFTWEKSPDIPDEDTSLTVFNQGNLMPSTKILTFY RKEPFNLEAKYATPEKLPGKTKPWIGNFSVKGVKADAKDDFMICKLKARLNLHGILNVDQG YYVEEVEVEEPIPEDPEKEPEVSFELNPFHVRSSSEERAVKRVRTGQPVLTSTKAMDIEDK PKTRKVKSQVRKGDLPISSATSGLDDSSKSASAEQEAAMIMEDKLVSDTEEKNELEAYIYE MRNKLDDQYSEFASEEEKLKLNERLTTTEDWLYDEGEDATKATYIAKMDEIRMLAGPITQR HFDKVEEERAAAQAISDAEAAKKRELAEAAKAAALAQEAAMSKDEVMAEADSSKNSEGFE ADSK