

Rothamsted Research Harpenden, Herts, AL5 2JQ

Telephone: +44 (0)1582 763133 Web: http://www.rothamsted.ac.uk/

Rothamsted Repository Download

A - Papers appearing in refereed journals

Mitchell, R. A. C. 2024. Identification of universal grass genes and estimates of their monocot-/ commelinid-/ grass-specificity. *Bioinformatics Advances.* p. vbaf079. https://doi.org/10.1093/bioadv/vbaf079

The publisher's version can be accessed at:

https://doi.org/10.1093/bioadv/vbaf079

The output can be accessed at:

https://repository.rothamsted.ac.uk/item/99007/identification-of-universal-grass-genesand-estimates-of-their-monocot-commelinid-grass-specificity.

© 7 April 2025, Please contact library@rothamsted.ac.uk for copyright queries.

14/04/2025 08:25

repository.rothamsted.ac.uk

library@rothamsted.ac.uk



Identification of universal grass genes and estimates of their monocot-/ commelinid-/ grass-specificity

Rowan A. Mitchell^{1, 2*}

¹Rothamsted Research, United Kingdom, ²Independent researcher, United Kingdom

Submitted to Journal: Frontiers in Plant Science

Specialty Section: Plant Bioinformatics

Article type: Original Research Article

Manuscript ID: 1396997

Received on: 06 Mar 2024

Revised on: 12 Apr 2024

Journal website link: www.frontiersin.org





Scope Statement

The manuscript describes a novel bioinformatics pipeline that aims to identify all universal protein-coding genes in grasses. It does this using genomes, gene models, and ortholog tables for 16 grass species from the Ensembl Plants major plant bioinformatics resource. Genes are organised into groups to optimise HHM profile scores and novel gene models are discovered in genomes using these profiles. Specificity to monocots / grasses is defined based on best hits to profiles from non-grass species. Users can access the results from supplementary tables and large datasets made available online.

Conflict of interest statement

The authors declare a potential conflict of interest and state it below

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision

CRediT Author Statement

Rowan Andrew Craig Mitchell: Conceptualization, Formal Analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing.

Keywords

Monocot, grass evolution, gene model, functional orthologs, Genomics

Abstract

Word count: 286

The evolutionary success of grasses is due to characteristics of resilience and fast growth in open habitats that led to their underpinning of agriculture and is attributable to many grassspecific traits. Genes responsible for these traits are likely specific to grasses, highly conserved and present in all grasses (universal genes) as they perform essential functions for fitness. A bioinformatics pipeline was developed to identify such genes using 16 grass full genomes in Ensembl Plants release 56. The first steps used existing gene models to generate groups of grass orthologs to rice and maize genes present in most grass species and refined membership of these groups such as to optimise the Hidden Markov Model (HMM) profile score from the HMMER package. These were then supplemented using new gene models found in grass genomes with the genBlastG tool; this step increased the number of universal groups by >2-fold to give 12,855 highly conserved, universal groups. Specificity for these groups was assessed using closest matching gene models from nonmonocot species. Possible cut-off values were tested with sets of known genes expected to be either of common function for all plants, or of commelinid-/ grass-specific function. A specificity metric based on HMM score from grass group profiles performed better than % identity as a means of discriminating between these common and specific function test sets. Using an appropriate cut-off for this metric, 5,701 of the groups were identified as monocot-/ commelinid-/ grass-specific of which 72% appeared to be grass specific. These results comprise the universal_grass_peps database available at DOI doi.org/10.23637/rothamsted.98ywz. This database can be searched by researchers to determine whether their experimentally identified grass genes match universal groups and, for those that do, to obtain systematic estimates of monocot-/ commelinid-/ grassspecificity.

Funding information

This work was made possible by funding from the UK Biotechnology and Biological Sciences Research Council awarded to RACM in grant BB/K007599/1 and to Rothamsted Research as strategic grant "Designing Future Wheat".

Funding statement

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article.

Ethics statements

Studies involving animal subjects

Generated Statement: No animal studies are presented in this manuscript.

Studies involving human subjects

Generated Statement: No human studies are presented in the manuscript.

Inclusion of identifiable human data

Generated Statement: No potentially identifiable images or data are presented in this study.



Data availability statement

Generated Statement: The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

- 1 Identification of universal grass genes and estimates of their
- 2 monocot- / commelinid- / grass-specificity.
- 3
- 4 Rowan A. C. Mitchell
- 5 rowan.mitchell@rothamsted.ac.uk
- 6 Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ, U.K.
- 7 https://www.rowanmitchell-grassscience.co.uk/
- 8
- 9

10 Abstract

11

12 The evolutionary success of grasses is due to characteristics of resilience and fast growth in open habitats that led to their underpinning of agriculture and is attributable to many grass-13 14 specific traits. Genes responsible for these traits are likely specific to grasses, highly conserved and present in all grasses (universal genes) as they perform essential functions 15 for fitness. A bioinformatics pipeline was developed to identify such genes using 16 grass full 16 genomes in Ensembl Plants release 56. The first steps used existing gene models to 17 generate groups of grass orthologs to rice and maize genes present in most grass species 18 and refined membership of these groups such as to optimise the Hidden Markov Model 19 (HMM) profile score from the HMMER package. These were then supplemented using new 20 21 gene models found in grass genomes with the genBlastG tool; this step increased the number of universal groups by >2-fold to give 12,855 highly conserved, universal groups. 22 Specificity for these groups was assessed using closest matching gene models from non-23 24 monocot species. Possible cut-off values were tested with sets of known genes expected to 25 be either of common function for all plants, or of commelinid- / grass-specific function. A 26 specificity metric based on HMM score from grass group profiles performed better than % 27 identity as a means of discriminating between these common and specific function test sets. 28 Using an appropriate cut-off for this metric, 5,701 of the groups were identified as monocot-/ 29 commelinid- / grass-specific of which 72% appeared to be grass specific. These results 30 comprise the universal_grass_peps database available at DOI 31 doi.org/10.23637/rothamsted.98ywz. This database can be searched by researchers to 32 determine whether their experimentally identified grass genes match universal groups and, 33 for those that do, to obtain systematic estimates of monocot- / commelinid- / grass-

35

34

specificity.

36 Introduction

37

38 Grasses (Poaceae) are of huge ecological importance, dominating open habitats in which 39 they played a fundamental role in forming (Jacobs et al., 1999;Kellogg, 2001) such that this 40 group of organisms now covers one third of global land area. Indeed it has been suggested that the rise of the grasses some 40 MYA was a key event in earth history, changing the 41 water cycle, carbon cycle and climate permanently (Retallack, 2001). Their adaptation to 42 43 open habitats has made them suited to adoption in agriculture and all the origins of human civilisation are associated with domestication of cereals and/or of grazing animals. Today, 44 about 70% of the calorie intake for humans comes directly or indirectly from grasses 45 (FAOSTAT, 2019). 46

Grasses co-evolved with large grazing mammals which few other plants can withstand 47 during early growth giving rise to the open grassland habitats (Stebbins, 1981). Key grass 48 adaptations to this ecosystem include: morphology that allows meristems to avoid 49 consumption and fire damage allowing regrowth; tissues rich in silica to resist herbivory and 50 51 stress (Mitani-Ueno and Ma, 2021); stomata that can respond faster than those of other 52 plants to rapidly changing conditions of open habitats (Chen et al., 2017); cell walls 53 containing ferulate implicated in lowering digestibility and stress resistance (Chandrakanth et al., 2023); unique inflorescence and seed characteristics for efficient reproduction (Kellogg, 54 55 2001). These traits are the result of specific protein-coding genes, non-coding genes and 56 regulatory genomic elements that arose in the evolution of grasses; here my aim was to develop a pipeline to identify the protein-coding genes (henceforth referred to as "genes" for 57 brevity) involved in grass-specific traits. 58

Relatively few genes involved in these traits have been demonstrated experimentally
(examples listed in Table 1). Among the best characterised are Lsi1, Lsi2 and Lsi6 genes
encoding silicic acid transporters that are required for Si accumulation and distribution. Both
monocots and dicots have homologs of Lsi1 and Lsi2 that transport silicic acid, but Lsi1

63 differs in polarity and localisation in monocots; monocots additionally have Lsi6 transporters 64 that direct Si transport within nodes (Ma and Yamaji, 2015; Mitani-Ueno and Ma, 2021). 65 Grass cell walls differ from those in dicots in several respects and genes responsible for the 66 grass-specific features are now known for: presence of (1,3;1,4)-beta-D-glucan (Burton et 67 al., 2006), ferulate moieties on the polysaccharide arabinoxylan that can cross-link xylan 68 chains or xylan to lignin (Feijao et al., 2022; Chandrakanth et al., 2023), lignin monomer tricin 69 (Lam et al., 2015) and beta-expansins which specifically mediate expansion of the differently 70 composed grass primary cell walls (Sampedro et al., 2015). Numerous monocot-specific 71 regulatory genes implicated in determining the unique morphology of grass inflorescence 72 have been identified; some of best characterised are the ramosa2 (Bortiri et al., 2006) and LOFSEP transcription factors (Kobayashi et al., 2012;Wu et al., 2018). Finally, some genes 73 74 involved in the fast-responding grass stomata such as guard cell SLAC1 anion channel have 75 been experimentally described (Schäfer et al., 2018).

76

I postulated that these grass genes and others responsible for functions specific to monocots 77 / grasses that are key to grass fitness will be (1) present in all grasses i.e. universal, (2) 78 highly conserved (3) have no close homologs in species outside monocots. The concept of 79 80 universality of genes – matching genes being present in all organisms within a taxonomic 81 unit - is a useful guide to their importance for fitness and implicitly groups genes by function 82 (Kriventseva et al., 2018). On point (3), it is convenient to consider monocot- and grass-83 specificity together because the large number of non-monocot plant genomes and wealth of 84 gene knowledge (particularly for Arabidopsis) make for a better reference set than the few, less studied non-grass monocot genomes. Also many key gene functions may have evolved 85 first in monocots and then been expanded by gene duplication in grasses. Thus the aim is to 86 capture those genes with key functional innovations that arose in monocots or grasses and 87 88 have not diverged further within the grasses; from typical estimates of timescales for origin of 89 monocots and divergence of grasses (Bouchenak-Khelladi, 2007), these function

90 innovations would have occurred in the period between 150 and 50 MYA.

91

92 The likelihood that two genes from different species share the same function increases with 93 the similarity of the encoded peptide sequences. For a given level of sequence similarity, it is 94 thought that they are more likely to share function if they are orthologs, i.e. descended from the same gene in the common ancestor (Gabaldón and Koonin, 2013). Here I also assumed 95 that universality of genes, i.e. if similar genes are found in every species of a taxonomic unit, 96 can also be taken as supporting common function as it may imply a role in a trait essential 97 98 for fitness. Furthermore, using this set of genes of putative common function allows the use of profiles that emphasise the conserved sequence elements that are key to that function 99 rather than weighting the whole sequence equally. 100

101

I used these principles to design the novel bioinformatics pipeline described here which aims 102 103 to: (a) identify a maximal set of groups of highly similar genes found in all grasses with each group having putative common function (b) assign estimates of how specific these functions 104 are to monocots / commelinid- / grass species based on closest hits from species outside 105 these taxa. Using the set of the genes described above to define a cut-off for specificity, 106 107 groups were classified as having monocot- / commelinid- / grass-specific or non-specific functions. I report some of the characteristics of these specific gene sets. Finally I discuss 108 109 uses and limitations of these predictions.

110

111

112 Methods

113

114 Predefined gene sets

115 To help test the pipeline output and to select cut-off values two sets of genes were pre-116 defined. The small number of monocot- /commelinid- / grass specific protein-coding genes of 117 known function (Table 1) were used as a specific test set. A list of proteins of known function expected to be common across all plants was also compiled to act as the non-specific test 118 119 set. This non-specific set was derived from ribosomal subunit proteins using RPG database 120 (Nakao et al., 2004) (http://ribosome.med.miyazaki-u.ac.jp) and enzymes or enzyme 121 subunits in amino acid synthesis, glycolysis, photosynthetic electron transport, CBH cycle and nucleotide synthesis from OryzaCyc database which had identical steps in AraCyc 122 database within Plant Metabolic Network (Hawkins et al., 2021) giving a total of 240 rice 123 peptides (Table S1). 124

125 Pipeline overview

Figure 1 shows a scheme of the pipeline which takes input data downloaded from Ensembl Plants, processes these using custom software and public packages and generates datasets that populate a novel database called universal_grass_peps.

The following input data were manually downloaded from the Ensembl Plants database 129 (Bolser et al., 2016) release 56 (https://feb2023-plants.ensembl.org/): peptide sequences 130 131 (peps) from gene models for 16 grass species and 58 non-grass species (Table S2), the full genome sequences of the grasses with their gene annotations, and the ortholog tables of 132 rice and maize gene models to all other grasses (downloaded using the Biomart tool). Rice 133 and maize were chosen as the reference species because they are intensively studied crops 134 with well annotated genomes representing respectively the BEP and PACMAD clades that 135 together include all grasses. 136

137 All operations on these input data were carried out on the Rothamsted Linux High

138 Performance Cluster using custom Perl (with Bioperl routines; Stajich et al., 2002) and bash

139 scripts to run bioinformatics tools and process data. These scripts are available at

140 <u>https://github.com/Rowan-ACM/universal_grass_peps</u>. The complete pipeline took 11 days

141 of run time on the cluster to complete.

The methods used in the different components of the pipeline shown in Fig. 1 are describedbelow.

144 Identification of highly conserved peptides present in all grasses (Find Universal Groups145 block in Fig. 1)

146 Using peps from gene models of the 16 grass species, any identical ones were removed but 147 all non-identical peps from splice variants were retained (for convenience, "gene" is used here to mean a unit encoding a unique peptide). For the rice and maize reference species, 148 using BLAST+ package (Camacho et al., 2009) a blastp search (parameters: -evalue 1.e-5 -149 max target seqs 50 -seq no -max hsps 1) of peps was conducted against all others within 150 151 same species and defined clusters where peps are >90% identical in both directions of a pairwise comparison for all comparisons within a cluster. Out of 40,196 rice peptides, 6% 152 were in clusters with >1 member, mostly highly similar splice variants. In maize 37% out of 153 62,559 peps were in such clusters; this higher percentage is expected in maize due to the 154 155 recent whole genome duplication (Swigonova et al., 2004) and greater propensity for tandem duplications (Guo et al., 2019). An ortholog table from the Ensembl multiple tables was 156 defined where each entry was defined by a primary key (group ID) of the rice peptide or 157 158 peps cluster ID. Where there was no maize ortholog, the most similar maize pep was found 159 with blastp and if this was not orthologous to another rice gene, added all the other grass genes orthologs of the maize gene to the group and group ID was set to composite of rice 160 and maize seed cluster IDs. Groups from maize were also allowed where there was no rice 161 ortholog or similar pep sequence (as this could be added by the later genBlastG step). Other 162

163 grass species orthologs to rice and/or maize were assigned exclusively to a single group 164 based on highest ranking by Ensembl ortholog confidence flag (0 or 1; defined from tree-165 compliance and, in a small number of cases, whole-genome alignment and gene order 166 conservation; https://plants.ensembl.org/info/genome/compara/peptide_compara.html) then 167 sequence similarity. Groups which had entries for fewer than 12 of the 16 grasses were 168 deleted and genes orthologous to remaining groups were reassigned according to this ranking. At this stage (Box 1 Fig. 1) there were groups of multiple genes per grass all of 169 170 which were classed orthologous to rice and/or maize. Using two reference species in this 171 way allows for similar non-orthologous genes of potentially common function to be grouped 172 together due to descending from two paralogs. But by using the ortholog information orthologous peps were more likely to be assigned to same group than non-orthologs with 173 same level of similarity. This is designed to help to group by function in accordance with the 174 175 principle that orthologs are more likely to share function at a given level of sequence similarity (Gabaldón and Koonin, 2013). 176

The next step (box 2 Fig. 1) was to optimise membership of groups keeping only one peptide 177 sequence per species. This approach makes the profile scores comparable across all 178 179 groups and avoids biasing profile to species with many members in a group. HMM profiles of each group were initially generated using the top ranked pep sequence for each species. To 180 181 make HMM profiles, all the group member peps were aligned using MUSCLE v3.8.1551 with 182 default parameters (Edgar, 2004) then the HMM profile was generated from this multiple alignment with hmmbuild (parameters --amino --fragthresh 0) and hmmpress commands 183 184 from HMMER package version 3.3.2, Nov 2020 (Eddy, 2022). Similarity scores of the member sequences against their own profile were obtained using hmmscan (all hmmscan 185 steps in pipeline used parameter E 1.e-7, other parameters default). To compare across 186 groups, this score was normalised to a maximum possible score obtained with the 187 188 consensus sequence of the profile (generated by hmmemit command) as the query to derive 189 a HMM relative score (R). Then group members were each substituted with all the

alternative peptide sequences for this group and species; if R was improved by > 0.01 the substitution was kept as the group member; this requirement means that peptide sequences ranked as best orthologs in previous step tended to be kept as group members. It was found that groups could be further improved by using grass Ensembl gene models hits to the HMM profile found with hmmscan that were not members of other groups; these are peps not found by previous steps probably because they were not in ortholog tables. Again, these peps were assigned as group members if they improved R by >0.01 (box 3 Fig. 1).

197 In the next step (box 4 in Fig. 1) the genBlastG tool was used (She et al., 2011) which searches for gene models with canonical splice junctions in genomic sequence using a 198 199 query peptide sequence; here the consensus from HMM profile for the group was used as 200 the query. For each group, and for each grass where the current member was missing or low 201 scoring, the relevant grass genome was searched with genBlastG (v138, parameters -p genblastg -v 2 -h 0 -j 3 -r 1 -norepair). Any hits discovered by genBlastG were checked that 202 203 they were novel by comparing exon coordinates with those of all Ensembl gene models using gff files. Using criteria as above, if a novel gene model from genBlastG improved the 204 profile, it was adopted as the group member for that species and the HMM profile was 205 206 rebuilt. A maximum of 4 genBlastG gene models were adopted so every profile has at least 207 12 Ensembl gene models. At the completion of this process, the R value was recalculated 208 for each member and groups where the lowest scoring member had R < 0.65 or had missing 209 members were discarded; the cut-off of 0.65 is a criterion for high conservation and the value was selected as that for which 90% of the pre-defined expected universal non-specific 210 211 genes (Table S1) groups passed. HMM profiles from the complete set of groups that pass these were compiled into a single HMMER database, the universal_grass_peps HMM 212 database. 213

214 Matches of grass genes to universal groups (box 5 in Fig. 1)

215 All scores for Ensembl grass peps against the universal_grass_peps HMM database for all 216 groups were obtained. All non-members that had scores of R > 0.65 to any group were 217 allocated as associate peptides allowing many-to-many relationships (this allows a lookup 218 search with any peptide ID as guery to find all groups to which a peptide is similar). To check 219 whether some universal_grass_peps groups can be regarded as likely same function, the R 220 of grass group members against other group HMMs were obtained. Where all members of a 221 group scored > 0.65 for another group and vice versa these groups were allocated to a 222 supergroup of potential same function.

223 Monocot- / Commelinid- / Grass- Specificity (Estimate Specificity block in Fig. 1)

Scores were obtained for the best-matching non grass peptide sequence from all the 58 non grass species against universal_grass_peps HMM database for all groups. A metric of specificity S for each group was evaluated, defined as R of the lowest scoring grass member of this group minus R of highest scoring non-monocot peptide. By definition a value of S <= 0 means the non-monocot peptide scores highly enough to be included so the group is completely non-specific.

Different cut-off values for this threshold were investigated using the groups containing the
genes from the pre-defined specific or non-specific test sets. For comparison of the S metric
with simple pairwise percentage identity, this was calculated from global alignment by
MUSCLE of the rice member of the group to its closest non-monocot hit identified by blastp.

234

235 Functional annotation of monocot- / commelinid- / grass-specific groups

236 To characterise the functions of the set of groups classified by the pipeline as monocot- /

237 commelinid- / grass-specific groups, functional annotations were obtained.

238 General gene descriptors and Gene Ontology terms from Ensembl Plants were assigned to

groups from their member rice and maize peps. Where present, linked publications, gene

- descriptors and symbols and trait ontology were assigned to groups from database entries
 for their member peps taken from RAP-DB (Sakai et al., 2013) and KnetMiner-rice for rice
 and MaizeMine (Shamimuzzaman et al., 2020) for maize and KnetMiner -wheat for wheat.
 Entries were retrieved from web interfaces except for KnetMiner where cereals knowledge
 graph (Hassani-Pak et al., 2021) with programmatic access was used to retrieve gene-TO
 and gene-GO relations for wheat and rice genes along with supporting publications.
- 246

247 Results

248

Identification of highly conserved peptides present in all grasses (Find Universal Groups
block in Fig. 1).

251 Initial steps (boxes 1-3 in Fig. 1) identified 17,816 groups of similar peps that were present in 252 at least 12 of the 16 grass species from their original gene models present in Ensembl 253 Plants release 56. Of these, 6,354 groups passed criteria for universality and high conservation (i.e. had members for all 16 species and minimum R > 0.65). However, correct 254 255 gene models are frequently missing from annotated genomes particularly where there is no 256 transcript information to support these as is often the case for lower expressed genes in less well studied species. Therefore the genomic sequences were searched for gene models for 257 each group and for each gene model that was missing or low scoring using the genBlastG 258 259 tool with consensus peptide sequence of the group HMM profile as query (box 4 in Fig. 1). 260 By incorporating the new gene models identified into groups the number of highly-conserved universal groups was more than doubled from 6,354 to 12,855 showing the importance of 261 the genBlastG step. The species break-down of the new gene models obtained by 262 genBlastG (Table 2) within these groups shows the newer genomes from Saccharum 263 264 spontaneum and Lolium perenne have the most whereas the intensively studied wheat with extensive transcript resources has the fewest. 265

266 The results for universal groups can be compared with those from the OrthoDB database which allows users to select for ortholog groups that are present in a minimum number of 267 species (Kriventseva et al., 2018). At the Poales level in OrthoDB release 10 there are 2,581 268 269 ortholog groups that are present in all 11 grass species and there is substantial overlap with the groups here with 85% of rice RABP IDs are present in universal groups before the filter 270 for high conservation. At this stage far more universal groups were recovered than from 271 272 OrthoDB and this seems to be mostly due to the genBlastG step rather than relying on 273 existing gene models.

274 The set of 12,855 highly conserved, universal groups obtained from the above steps are 275 here termed the universal grass peps database. These groups contain sequences that all 276 match the profile well but also contain different degrees of divergence. Two example multiple 277 alignments used to generate the HMMs for two groups are shown in Figure 2. These show 278 high conservation including for the novel genBlastG gene models but also reveal some of 279 the inherent complexities found in most profiles; group Os03t0786600-01 has 280 overwhelmingly similar sequences but also has some signs of divergence at the C-terminal 281 between BEP (species 1-8) and PACMAD clade grasses (species 9-16), and group 282 Os02t0763000-01 has a section found only in one species. Nevertheless these alignments 283 do support the hypothesis of highly similar function common to all grasses for these groups.

284

285 Matches of grass genes to universal groups (box 5 in Fig. 1)

All grass gene model peps were searched against the HMM profiles of 286 universal_grass_peps for hits with R above the cut-off of 0.65; if these are not the member of 287 288 any group they are classified as associated to the group. The total number of associated peps for each species is shown in Table 2 and generally reflects the degree of gene 289 290 duplication. The grass pep hits are also used to define supergroups; if all members of one 291 group are hits above cut-off to another group, the two groups are assigned to super-groups 292 of closely related function. A total of 799 supergroups were identified (Table S3). 293 Supergroups can contain groups with same molecular function but differing regulation due to 294 sub-functionalisation.

295

296 Monocot- / Commelinid- / Grass- Specificity (Estimate Specificity block in Fig. 1)

All peps from the 58 non-grass species in Ensembl Plants were scored against the HMM profiles of universal_grass_peps. The results were used to derive the specificity metric S for each group, where S is minimum R value from group members minus maximum R value for
any peptide from non-grass to give monocot-/ commelinid-/ grass-specificity. Distinguishing
between monocot-specificity, commelinid-specificity and grass-specificity is dependent on
maximum R values from only 3 species (two non-grass commelinids and one noncommelinid monocot) so these sub-classifications are less secure, and the overall monocot-/
commelinid-/ grass-specificity is emphasised here.

The S metric is a measure of sequence divergence from the grass profile that can be used 305 306 as a basis for an initial hypothesis of function divergence in the same way that other 307 sequence-based measures are used. The pre-defined test sets were used to gauge the 308 performance of S as a means of determining specificity, i.e. the non-specific test set of 215 309 peps expected to have common function in all plants because the fundamental processes 310 they are responsible for are not thought to have diverged (Table S1) and the specific test set 311 of 16 peps with monocot- / commelinid-/ grass-specific functions (Table 1). The S metric was 312 compared with simple pairwise % identity with the best non-monocot hit for these sets (Figure 3); S performs better than % identity at discriminating between the two sets as 313 choosing highest cut-off with no false negatives gives 11.6% false positives using S and 314 315 14.4% false positives using pairwise percentage identity. Using sequence similarity (e.g. 316 from BLOSUM62) rather than identity did not improve performance of pairwise alignment as a measure (data not shown). 317

Applying the cut-off S of >0.25 which gave 11.6% false positive and no false negatives with the test sets (Figure 3) to the complete set of 12,855 groups gave 5,701 defined as monocot- / commelinid- / grass-specific. This set was divided into subsets classified as probably monocot-specific (355 profiles), commelinid-specific (1,260 profiles) and grassspecific (4,086 profiles) based on values of S calculated from best hits for each taxonomic level and is listed in Table S4.

324

325 Functional annotation of monocot- / commelinid- / grass-specific groups

Functional annotations for these 5,701 specific groups were derived from public annotations 326 of their rice, wheat, and maize members. Most (~90%) have no linked publications and only 327 general descriptors and high-level GO terms based on domains. When the set is ranked by 328 329 S metric, the groups with least similarity to any non-grass pep often have nothing known but 330 a prominent domain is "Cyclin-like F-box domain" which occurs in 4 of the 20 most grassspecific profiles (Table S4). Proteins containing this domain were also highlighted in an early 331 study attempting to identify grass-specific proteins (Campbell et al., 2007). F-box domains 332 are associated with protein-protein interaction e.g. for the regulation of other proteins by 333 334 ubiquitination.

A wider view of the processes in which the monocot- / commelinid- / grass-specific genes 335 take part can be gained from analysis of GO terms assigned in RAP-DB and MaizeMine, 336 based mostly on recognition of domains and functions of homologs. Of all the biological 337 process GO annotations, most are assigned to at least one group suggesting there are some 338 339 monocot- / commelinid- / grass-specific aspects of most processes in grasses. The 340 processes that are dominated by these specific genes are shown by the terms which are enriched; there is clear enrichment of groups of regulatory proteins, especially those 341 342 involved in control of transcription and of protein activity (Table 3). Some specific enriched 343 terms include ones that might be expected such as xylan biosynthesis and leaf development 344 but also include fundamental processes such as cell cycle. Enriched molecular function GO terms are mostly DNA-binding and enzyme activities; the most enriched enzyme category is 345 hydroxycinnamoyl transferase activity (Table 3) which may reflect the importance of these 346 347 moities on lignin and xylan polymers in grass cell walls (Chandrakanth et al., 2023).

For the minority of groups with associated publications, the publications, gene descriptors and trait ontology (TO) terms from the RAP-DB, MaizeMine and KnetMiner databases were assigned. The traits defined by TO terms are associated with variants of the member rice,

- maize and/or wheat genes from evidence in the publications. For the monocot- / commelinid/ grass-specific groups (Table S4), particularly common traits affected are grain size (90
 groups), flowering time (62 groups), with numerous morphology traits as might be expected.
- However also common are traits for insect / pathogen defence and abiotic stress resistance.
- 355 The complete set of 5,701 groups defined as monocot- / commelinid- / grass-specific
- together with specificity estimates and all functional annotation are in Table S4.

357

358 Discussion

359

Genes of common function that occur in all species within a taxonomic unit (universality) 360 indicate that the function is likely key to fitness. Although sequence similarity is a measure of 361 362 likelihood of shared function, using existing bioinformatic resources it is not straightforward to compare genes in a systematic way, nor to check for criterion of universality given 363 variation in completeness of genome annotation. The new approach described here provides 364 365 predictions of all universal grass genes with putative common function and estimates of their 366 specificity to monocots / commelinids / grasses. It should be noted that since the pipeline generates groups with putative common function that can contain any similar gene, not just 367 true orthologs, it is not directly comparable to software like OrthoFinder that identify 368 orthologous groups (Emms and Kelly, 2019). Rather, ortholog tables are an input to the 369 370 pipeline as a starting point for seeding groups (box 1 in Fig. 1) but these predicted orthologs can be replaced in a later step by alternative peps from the same species if they match the 371 profile better (box 3 in Fig. 1). A novel aspect of the pipeline is the emphasis on universality 372 373 which led to the incorporation of the genBlastG step to find missing genes (box 4 in Fig. 1); this step generated 14,038 new gene models. The fact that grass species like wheat that 374 have more RNAseg data require fewer of these gene models (Table 2) suggests that future 375 grass RNAseg studies will validate many of them. The use of a metric based on HMM profile 376 score to estimate how specific the function of a universal group is to monocot / commelinid / 377 378 grass species is another novel aspect of the pipeline; it provides a systematic basis for an assertion of such specificity for genes of unknown function. 379

380

381 Importance of monocot- / commelinid- / grass-specific genes

Grasses typically have a haploid set of about 40,000 protein-coding genes. The analysis here indicates that about 12,000 of these are universal in grasses and that about half of universal genes are monocot- / commelinid- / grass-specific. These genes are enriched for 385 regulatory functions (Table 3) as might be expected given the radically different organisation 386 and morphology of grasses. The great majority of genes in the specific sets are of unknown 387 function which reflects our lack of understanding of molecular mechanisms underlying grass-388 specific characteristics. However, the GO term annotation indicates that these genes are 389 likely involved in virtually every process in grasses and are particularly dominant in the 390 enriched ones shown in Table 3 which include cell wall processes and stomatal regulation as 391 might be predicted but also some less expected such as control of epigenetic marks and 392 chloroplast movement.

393 The importance of variants of the monocot- / commelinid- / grass-specific genes for crop 394 traits is seen from publications associated to the identified sets (Table S4) including 395 numerous variants associated with grain yield, abiotic stress and defence. Where a trait is 396 known to be commelinid- or grass-specific, the classifications generated here can help to 397 identify candidate genes involved in the trait. In our own work on dietary fibre QTLs in wheat 398 grain, candidate genes identified as likely commelinid- / grass-specific were prioritised as 399 dietary fibre is mostly feruloylated arabinoxylan (AX) that only occurs in commelinid species. 400 The causal allele was eventually shown to be a variant of one such gene – a commelind-401 specific peroxidase involved in cross-linking AX (Mitchell et al., 2023). There must be many 402 more valuable natural and induced variants of these genes yet to be discovered and the 403 classifications generated here could help in candidate identification.

404

405 Limitations of approach

All high-throughput predictions of shared function based almost entirely on peptide sequence need to be used with caution and cannot substitute for detailed knowledge of the particular protein. The approach here should be treated as a first best guess of shared function similar to comparing percentage identity (as biologists often do as a first step) but more likely to be accurate (Fig. 3) as the HMM approach weights the conserved parts of sequence important for function, exploiting the fact that the identified genes are highly conserved and present in 412 all grasses. The gene groups would be expected to include nearly all cases of genes which 413 have identical function in all grasses, but they can also include cases where there are highly 414 similar functions with divergent aspects. This is because non-conserved regions do not 415 affect the profile score much so if there is a conserved core and divergent, species-specific 416 functional aspects of the sequence they can still pass the highly conserved filter. Therefore 417 the next step after identifying a group of interest based on its S score should be to inspect 418 the multiple alignment (as in Fig. 2) to judge the extent of divergence in different grasses; all 419 12,855 multiple sequence alignment files are available in the universal_grass_peps 420 database.

Too much divergence from the group profile will lead to the group being excluded. These cases will likely include genes that were important in evolutionary history of grasses but have subsequently diverged in adaptation to the many different ecosystems that grasses occupy since their divergence some 55 MYA (Bouchenak-Khelladi, 2007) including the major bifurcation into the BEP and PACMAD clades with respectively C3 and C4 photosynthesis.

426

427 Uses of universal_grass_peps database

Where experiments reveal large sets of grass genes or peps such as transcriptomics, 428 proteomics or genes underlying QTLs, they are inevitably dominated by genes with little or 429 430 no information on function. Even for rice, probably the most studied grass, only 13% of genes in RAP-DB database (Sakai et al., 2013) have associated publications and only a 431 432 minority of these publications specify function. For such unknown genes it is useful to have a systematic approach to identifying those that are of grass- / commelinid- / monocot- specific 433 434 function as this information can point to the nature of the process they are likely involved in. 435 For example a network of co-regulated genes identified from transcriptomics enriched for grass-specific functions indicates involvement of the network in a grass-specific trait such as 436 inflorescence development, Si deposition etc. Using the look-up tables generated, any set of 437

grass genes from the grass genomes used here can be used to find all those in, or
associated to, the universal groups and their categorisations as likely monocot-, commelinidor grass- specific. For other grass genes, the HMMs database for the universal groups can
be searched using the HMMER package. For genes with matches in the universal groups,
the value of the S metric is a measure of how different the group is from any non-grass pep
and the supplied multiple alignments can be used to judge divergence from the profile.

444 The universal_grass_peps database is available at

445 <u>https://doi.org/10.23637/rothamsted.98ywz</u>. On the top directory there is a user guide and

summary spreadsheet of all 12,855 groups; the HMM database, multiple sequence

- alignments, genBlastG gene models and lookup tables for grass genes are in subdirectories.
- 448

449 Future developments

The pipeline reported here is a first attempt to implement the concept of using universal 450 genes to identify groups of putative common function but could be improved upon with 451 452 different software in future. Improvements might be made by using recently released alternative packages for finding orthologs (Emms and Kelly, 2019) as the first step (box 1 in 453 Fig. 1) and gene models in genomes (Li, 2023) (to replace genBlastG for box 4 in Fig. 1) 454 455 with reportedly better performance. Further in the future, two more major changes would be 456 to use structural prediction and incorporate expression patterns. The use of HMMs is a 457 convenient and fast way of obtaining profiles for groups against which other sequences can be scored for matches but here it is actually a proxy for comparison of structures. A direct 458 comparison of predicted structures such as that generated by AlphaFold might be a better 459 approach. Also similar expression patterns of peptides from different species are a strong 460 461 indicator of shared function and would help resolve cases of sub-functionalisation (Das et al., 2016). However, I am not aware of any software packages capable of conveniently and 462

- 463 quantitatively comparing predicted structures or expression patterns that could be used to achieve these improvements in the pipeline at present. 464
- 465

Conclusion 466

A novel bioinformatics approach was used to try to identify all universal grass genes coding 467

proteins responsible for monocot- / commelinid- / grass-specific traits, making the first 468

469 estimates of the size of these sets. As part of this, 14,038 new gene models were generated

for 16 grass genomes. The resulting classifications of grass genes can help interpretation of 470

experimentally identified sets of grass genes and represent numerous gene research targets 471

to improve our understanding of grass-specific mechanisms. 472 review

473

474

Tables

T	ables					
Т	able 1 Known monoco	ot- /commelinid-/ grass-s	pecific genes.			
	trait	gene family	gene name(s)	reference species gene ID(s)	description	reference
	secondary metabolite, cell wall	cytochrome P450	CYP93G1; CYP75B4	Os04g0101400; Os10g0317900	production of tricin a secondary metabolite and lignin monomer specific to grasses / monocots	Lam et al., 2015
	cell wall	cellulose synthase-like F	OsCsIF2	Os07g0552800	makes (1,3;1,4)-beta-glucan, a component of grass cell walls absent in dicots implicated in formation of feruloyl-arabinfuranosyl	Burton et al., 2006
	cell wall	BAHD acyl-CoA transferases	BAHD01 / AT9; BAHD05 / AT1	Os01g0185300; Os01g0615300	precursor prior to additon to xylan, a key feature of commelinid cell walls mediates addition of hydroxycinnamates to	Chandrakanth et al., 2023
	cell wall	BAHD acyl-CoA transferases glycosyl transferase	PMT1; FMT / AT5	Os05g0136900; Os05g0278500	monolignols leading to commelinid-specific features on lignin implicated in addition of feruloyl-arabinofuranose to	Chandrakanth et al., 2023
	cell wall	family 61	XAX1	Os02g0329800	xylan, a key feature of grass cell walls	Feijao et al., 2022
	cell wall inflorescence	expansins MADS transcription	EXPB9 OsMADS1;	Os10g0548600 Os03g0215400;	grass-specific beta-expansins evolved to mediate expansion in grass primary cell walls	Sampedro et al., 2015
	morphology inflorescence	factor MADS transcription	OsMADS5	Os06g0162800	specifies spikelet identity in rice inflorescence PAP2 / OsMADS34 regulator of spikelet identity.	Wu et al., 2018 Kobayashi et al.,
	morphology	factor LOB domain	MADS34/PAP2	Os03g0753100	Controls developmental processes unique to grasses	2012
	inflorescence morphology	transcription factor MIP/aquaporin	ramosa2	Zm00001eb123060	ramosa2 responsible for genetic control of grass- specific inflorescence	Bortiri et al., 2006 Ma and Yamaji,
	Si transport	membrane proteins	LSi1	Os02g0745100	transporter required for active uptake of Si in grasses	2015
	Si transport	MIP/aquaporin membrane proteins S-type anion channel	LSi6	Os06g0228200	transporter required to control distribution of Si to leaves and panicle in rice contains a monocot-specific motif that confers nitrate-	Mitani-Ueno and Ma, 2022
	stomata	family	SLAC1	Os04g0574700	sensitivity to guard cell anion channel	Schäfer et al., 2018



grass species	Group members*	Members that are genBlastG gene models	Groups with associate peps	Max associate peps in one group	Total associate peps
Brachypodium_distachyon	12,855	312	4,776	25	9,926
Hordeum_vulgare	12,855	899	2,934	33	6,185
Leersia_perrieri	12,855	1,317	4,375	15	8,128
Lolium_perenne	12,855	2,142	3,084	28	6,059
Oryza_rufipogon	12,855	921	4,278	16	8,151
Oryza_sativa	12,855	1,819	3,060	16	4,912
Secale_cereale	12,855	491	2,810	39	7,977
Triticum_aestivum	12,855	178	10,965	91	68,522
Echinochloa_crus-galli	12,855	287	10,752	48	39,486
Eragrostis_curvula	12,855	1,329	5,026	27	9,830
Panicum_hallii_HAL2	12,855	212	4,166	17	7,957
Saccharum_spontaneum	12,855	2,305	7,419	36	16,944
Setaria_italica	12,855	796	4,199	23	8,085
Setaria_viridis	12,855	131	4,886	22	10,183
Sorghum_bicolor	12,855	276	4,485	36	9,167
Zea_mays	12,855	623	7,854	23	22,744

Table 2 Counts of peps or groups for each grass species in universal_grass_peps database

*by definition, all grass species have same number of group members

Table 3 GO terms that are enriched in Ensembl annotation for rice and maize members of the monocot-/ commelinid-/ grass-specific universal groups. All GO term names that occur in at least 5 groups and are enriched relative to rice, maize peps and to universal peps >2-fold are shown.

	number of monocot- /commelinid- /grass-specific groups	enrichment relative to Os, Zm peps	enrichment relative to all grass_universal peps
GO Domain: biological_process			
positive regulation of DNA-templated transcription	62	2.5	2.3
cell differentiation	48	2.9	2.1
negative regulation of catalytic activity	27	2.8	2.2
regulation of cyclin-dependent protein serine/threonine kinase			
activity	18	2.5	2.0
response to endoplasmic reticulum stress	13	2.7	2.3
regulation of jasmonic acid mediated signaling pathway	12	2.6	2.7
interstrand cross-link repair	12	3.5	2.4
cellular response to nitrate	12	5.9	2.5
mitotic cell cycle phase transition	12	2.4	2.4
nuclear-transcribed mRNA catabolic process, nonsense-mediated			
decay	12	2.1	2.1
positive regulation of transcription from RNA polymerase II			
promoter in response to heat stress	12	8.1	2.7
DNA-templated transcription termination	12	2.2	2.4
regulation of primary metabolic process	11	2.2	2.5
negative regulation of endopeptidase activity	11	2.2	2.7
xylan biosynthetic process	11	2.9	2.0
regulation of nitrogen compound metabolic process	11	5.8	2.5
gene silencing by RNA-directed DNA methylation	10	2.8	2.9
regulation of leaf development	10	6.0	2.1
plastid transcription	7	3.3	2.3
mRNA destabilization	7	3.3	2.9
purine nucleoside transmembrane transport	6	2.8	2.2
nuclear-transcribed mRNA catabolic process, exonucleolytic, 3'-5'	6	3.1	2.2
rRNA methylation	6	2.5	2.2
positive regulation of helicase activity	6	5.1	2.2
mitotic spindle assembly	5	3.2	2.1
male meiosis II	5	4.7	2.9
negative regulation of organ growth	5	3.1	2.9
positive regulation of defense response to bacterium	5	7.7	2.1
mitochondrial mRNA modification	5	2.2	2.1
piecemeal microautophagy of the nucleus	5	3.2	2.9
	_		_

positive regulation of mitochondrial translation	5	5.6	2.5
asymmetric cell division	5	3.0	2.9
chloroplast avoidance movement	5	2.2	2.9
malate transport	5	2.9	2.1
chloroplast accumulation movement	5	2.1	2.5
post-transcriptional regulation of gene expression	5	2.0	2.1
protein localization	5	2.3	2.1
regulation of mitotic cell cycle	5	2.6	2.1
response to red or far red light	5	2.7	2.9
GO Domain: cellular_component			
chromosome, telomeric region	8	2.4	2.1
RNA polymerase II transcription regulator complex	8	4.1	2.9
nuclear microtubule	6	2.2	2.9
cell periphery	6	3.1	2.2
plastid-encoded plastid RNA polymerase complex	5	3.5	2.2
DNA polymerase III complex	5	3.0	2.9
histone acetyltransferase complex	5	2.4	2.5
GO Domain: molecular_function			
sequence-specific DNA binding	197	2.4	2.3
DNA-binding transcription factor activity, RNA polymerase II-	70	2.2	
specific	79	2.2	2.2
RNA polymerase II cis-regulatory region sequence-specific DNA	70	2.5	2.2
binding	73	2.5	2.2
hydroxycinnamoyltransferase activity	30	4.5	2.2
DNA-binding transcription activator activity, RNA polymerase II-	20	2.2	2 5
specific	30	3.3	2.5
enzyme inhibitor activity	22	2.4	2.5
quercetin 7-O-glucosyltransferase activity	21	2.6	2.1
quercetin 3-O-glucosyltransferase activity	21	2.6	2.1
pentosyltransferase activity	14	2.4	2.1
ubiquitin conjugating enzyme binding	12	3.1	2.1
pectinesterase inhibitor activity	11	3.0	2.4
DNA-binding transcription repressor activity, RNA polymerase II-			
specific	11	10.6	2.4
histone acetyltransferase activity	10	3.0	2.2
glutathione binding	9	3.4	2.3
histone methyltransferase activity	8	4.5	2.1
ribonuclease activity	8	2.2	2.3
5'-3' exodeoxyribonuclease activity	7	5.2	2.0
myosin XI tail binding	7	2.8	2.9
galactosylgalactosylxylosylprotein 3-beta-glucuronosyltransferase			
activity	7	2.2	2.9
endo-1,4-beta-xylanase activity	7	3.7	2.0
double-stranded RNA binding	7	2.5	2.5
purine nucleoside transmembrane transporter activity	6	2.9	2.1
xylosyltransferase activity	6	3.4	2.5
i i 'I	-	-	

electron transporter, transferring electrons within the cyclic			
electron transport pathway of photosynthesis activity	5	2.1	2.9
strictosidine synthase activity	5	2.2	2.0
sucrose transmembrane transporter activity	5	6.4	2.0
myosin binding	5	2.8	2.9
ionotropic glutamate receptor activity	5	3.9	2.0
RNA-DNA hybrid ribonuclease activity	5	3.2	2.4
NAD+ ADP-ribosyltransferase activity	5	2.1	2.0
histone H3-methyl-lysine-9 demethylase activity	5	2.6	2.9
translation activator activity	5	6.4	2.4
telomeric DNA binding	5	3.2	2.4

References

- Bolser, D., Staines, D.M., Pritchard, E., and Kersey, P. (2016). "Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data," in *Plant Bioinformatics: Methods* and Protocols, ed. D. Edwards. (New York, NY: Springer New York), 115-140.
- Bortiri, E., Chuck, G., Vollbrecht, E., Rocheford, T., Martienssen, R., and Hake, S. (2006). ramosa2 Encodes a LATERAL ORGAN BOUNDARY Domain Protein That Determines the Fate of Stem Cells in Branch Meristems of Maize. *The Plant Cell* 18, 574-585.
- Bouchenak-Khelladi, Y. (2007). *Grass-evolution and Diversification: A Phylogenetic Approach.* PhD, Trinity College Dublin.
- Burton, R.A., Wilson, S.M., Hrmova, M., Harvey, A.J., Shirley, N.J., Stone, B.A., Newbigin, E.J., Bacic, A., and Fincher, G.B. (2006). Cellulose synthase-like CsIF genes mediate the synthesis of cell wall (1,3;1,4)-beta-D-glucans. *Science* 311, 1940-1942.
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. *BMC bioinformatics* 10, 1-9.
- Campbell, M.A., Zhu, W., Jiang, N., Lin, H., Ouyang, S., Childs, K.L., Haas, B.J., Hamilton, J.P., and Buell, C.R. (2007). Identification and Characterization of Lineage-Specific Genes within the Poaceae. *Plant Physiology* 145, 1311-1322.
- Chandrakanth, N.N., Zhang, C., Freeman, J., De Souza, W.R., Bartley, L.E., and Mitchell, R.A.C. (2023). Modification of plant cell walls with hydroxycinnamic acids by BAHD acyltransferases. *Frontiers in Plant Science* 13, doi.org/10.3389/fpls.2022.1088879.
- Das, M., Haberer, G., Panda, A., Das Laha, S., Ghosh, T.C., and Schäffner, A.R. (2016). Expression Pattern Similarities Support the Prediction of Orthologs Retaining Common Functions after Gene Duplication Events *Plant Physiology* 171, 2343-2357.
- Eddy, S. (2022). HMMER User's Guide: Biological Sequence Analysis Using Profile Hidden Markov Models. http://eddylab.org/software/hmmer/Userguide.pdf [Online]. Available: http://eddylab.org/software/hmmer/Userguide.pdf [Accessed].
- Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32, 1792-1797.
- Emms, D.M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* 20, 238.
- Feijao, C., Morreel, K., Anders, N., Tryfona, T., Busse-Wicher, M., Kotake, T., Boerjan, W., and Dupree, P. (2022). Hydroxycinnamic acid-modified xylan side chains and their cross-linking products in rice cell walls are reduced in the Xylosyl arabinosyl substitution of xylan 1 mutant. *The Plant Journal* 109, 1152-1167.
- Gabaldón, T., and Koonin, E.V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics* 14, 360-366.
- Guo, H., Jiao, Y., Tan, X., Wang, X., Huang, X., Jin, H., and Paterson, A.H. (2019). Gene duplication and genetic innovation in cereal genomes. *Genome Res* 29, 261-269.
- Hassani-Pak, K., Singh, A., Brandizi, M., Hearnshaw, J., Parsons, J.D., Amberkar, S., Phillips, A.L., Doonan, J.H., and Rawlings, C. (2021). KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnology Journal* 19, 1670-1678.
- Hawkins, C., Ginzburg, D., Zhao, K., Dwyer, W., Xue, B., Xu, A., Rice, S., Cole, B., Paley, S., and Karp, P. (2021). Plant Metabolic Network 15: A resource of genome-wide metabolism databases for 126 plants and algae. *Journal of integrative plant biology* 63, 1888-1905.
- Jacobs, B.F., Kingston, J.D., and Jacobs, L.L. (1999). The origin of grass-dominated ecosystems. *Annals of the Missouri Botanical Garden*, 590-643.
- Kellogg, E.A. (2001). Evolutionary history of the grasses. Plant Physiology 125, 1198-1205.

- Kobayashi, K., Yasuno, N., Sato, Y., Yoda, M., Yamazaki, R., Kimizu, M., Yoshida, H., Nagamura, Y., and Kyozuka, J. (2012). Inflorescence Meristem Identity in Rice Is Specified by Overlapping Functions of Three AP1/FUL-Like MADS Box Genes and PAP2, a SEPALLATA MADS Box Gene *The Plant Cell* 24, 1848-1859.
- Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A., and Zdobnov, E.M. (2018). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* 47, D807-D811.
- Lam, P.Y., Liu, H., and Lo, C. (2015). Completion of Tricin Biosynthesis Pathway in Rice: Cytochrome P450 75B4 Is a Unique Chrysoeriol 5'-Hydroxylase. *Plant Physiology* 168, 1527-1536.
- Li, H. (2023). Protein-to-genome alignment with miniprot. *Bioinformatics* 39, btad014.
- Ma, J.F., and Yamaji, N. (2015). A cooperative system of silicon transport in plants. *Trends in Plant Science* 20, 435-442.
- Mitani-Ueno, N., and Ma, J.F. (2021). Linking transport system of silicon with its accumulation in different plant species. *Soil Science and Plant Nutrition* 67, 10-17.
- Mitchell, R.a.C., Oszvald, M., Pellny, T.K., Freeman, J., Halsey, K., Sparks, C.A., Huttly, A., Specel, S., Leverington-Waite, M., Griffiths, S., Shewry, P.R., and Lovegrove, A. (2023). A high solublefibre allele in wheat encodes a defective cell wall peroxidase responsible for dimerization of ferulate moieties on arabinoxylan. *bioRxiv*, 2023.2003.2008.531735.
- Nakao, A., Yoshihama, M., and Kenmochi, N. (2004). RPG: the ribosomal protein gene database. *Nucleic acids research* 32, D168-D170.
- Retallack, G.J. (2001). Cenozoic expansion of grasslands and climatic cooling. *The Journal of Geology* 109, 407-426.
- Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., Wakimoto, H., Yang, C.C., Iwamoto, M., Abe, T., Yamada, Y., Muto, A., Inokuchi, H., Ikemura, T., Matsumoto, T., Sasaki, T., and Itoh, T. (2013). Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol* 54, e6.
- Sampedro, J., Guttman, M., Li, L.-C., and Cosgrove, D.J. (2015). Evolutionary divergence of β– expansin structure and function in grasses parallels emergence of distinctive primary cell wall traits. *The Plant Journal* 81, 108-120.
- Schäfer, N., Maierhofer, T., Herrmann, J., Jørgensen, M.E., Lind, C., Von Meyer, K., Lautner, S.,
 Fromm, J., Felder, M., Hetherington, A.M., Ache, P., Geiger, D., and Hedrich, R. (2018). A
 Tandem Amino Acid Residue Motif in Guard Cell SLAC1 Anion Channel of Grasses Allows for
 the Control of Stomatal Aperture by Nitrate. *Current Biology* 28, 1370-1379.e1375.
- Shamimuzzaman, M., Gardiner, J.M., Walsh, A.T., Triant, D.A., Le Tourneau, J.J., Tayal, A., Unni, D.R., Nguyen, H.N., Portwood, J.L., and Cannon, E.K. (2020). MaizeMine: a data mining warehouse for the maize genetics and genomics database. *Frontiers in Plant Science* **11**, 592730.
- She, R., Chu, J.S.-C., Uyar, B., Wang, J., Wang, K., and Chen, N. (2011). genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* 27, 2141-2143.
- Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert,
 J.G.R., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock,
 M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., and Birney, E. (2002).
 The bioperl toolkit: Perl modules for the life sciences. *Genome Research* 12, 1611-1618.
- Stebbins, G.L. (1981). Coevolution of grasses and herbivores. *Annals of the Missouri Botanical Garden*, 75-86.
- Swigonova, Z., Lai, J., Ma, J., Ramakrishna, W., Llaca, V., Bennetzen, J.L., and Messing, J. (2004). On the tetraploid origin of the maize genome. *Comparative and functional genomics* **5**, 281-284.
- Wu, D., Liang, W., Zhu, W., Chen, M., Ferrándiz, C., Burton, R.A., Dreni, L., and Zhang, D. (2018). Loss of LOFSEP transcription factor function converts spikelet to leaf-like structures in rice. *Plant Physiology* 176, 1646-1664.

Acknowledgments and Funding

My thanks to Keywan Hassani-Pak, head of bioinformatics at Rothamsted Research, for useful discussions and providing KnetMiner data, and to Melina Velasquez for assistance with data repository. This work was made possible by funding from the UK Biotechnology and Biological Sciences Research Council awarded to RACM in grant BB/K007599/1 and to Rothamsted Research as strategic grant "Designing Future Wheat".

Figure Legends

Figure 1 Pipeline that generates the database of highly conserved universal grass proteincoding genes and estimates of their monocot- / commelinid- / grass-specificity (universal_grass_peps). All the input data is taken from Ensembl Plants release 56 and the processing steps are carried out by custom scripts, using the external tools shown in blue text, to generate universal_grass_peps database.

Figure 2 Two example group multiple alignments from the universal_grass_peps set of groups. Sequences are from grass spp 1. *Brachypodium distachyon* 2. *Hordeum vulgare* 3. *Leersia perrieri* 4. *Lolium perenne* 5. *Oryza rufipogon* 6. *Oryza sativa* 7. *Secale cereale* 8. *Triticum aestivum* 9. *Echinochloa crus-galli* 10. *Eragrostis curvula* 11. *Panicum hallii* HAL2 12. *Saccharum spontaneum* 13. *Setaria italica* 14. *Setaria viridis* 15. *Sorghum bicolor* 16. *Zea mays*. Sequences predicted by genBlastG have names starting "genblast" others are Ensembl gene models. Max score is the score of the consensus against the HMM profile generated from the alignment.

Figure 3 Proportion of groups of pre-defined genes expected to be of non-specific function (blue line) or specific function for commelinid / grass species (red line) that pass varying cutoff thresholds for two metrics of specificity. Upper panel: percentage identity of closest nonmonocot hit to rice member of group. Lower panel: S metric defined as lowest HMM relative score of group member minus the top relative score for a non-monocot hit. In both panels the limit which gives no false negatives and minimum false positives is shown.

31

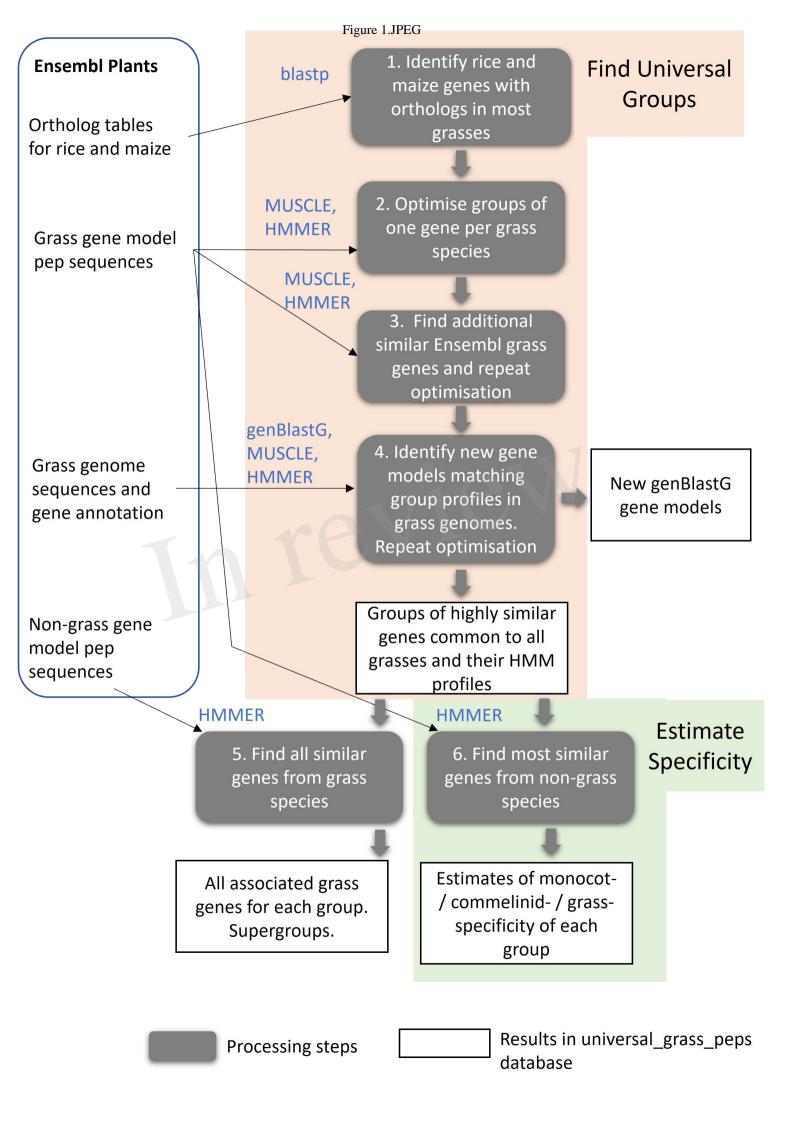
Supplementary Data

Table S1. Pre-defined set of grass genes expected to be universal and of non-specific function.

Table S2. All plant genomes used from Ensembl Plants release 56.

Table S3. Supergroups: contains groups which are so similar that all members would pass cut-off to be in another group within supergroup.

Table S4. Set of 5,701 groups classified as monocot- / commelinid- / grass-specific with details of specificity estimates, group properties, members, and functional annotation.



Group: Os03t0786600-01 max_score: 856

	1 10	20	30	40	50	60	70	80 90	100
1. KOK86779	MQSEAK - PRALEQUHP			ASEDGVPVI	SHARDIRPLAS	ARESPSPT	PMSAVSKLORS		REPAAEVESVGPO
2. HORVU.MOREX.r3.5HG0516920.1.CDS1	MRSLARPPRPPKPLPL								REPAAEVESVGPE
3. LPERR03G30940.1	MRHLAMRRREARRLA-				SHAGDLRPLVS				REPAAEVESVGPE
4. cds.KYUSt chr4.7551	MRSLAK-ARPPKPLCP	AAEVOSMGYV		ASEDGVPVI	AHARDIRPLAA	ARLISPSPT	PVSAVSKIRRS	ETSDRRVAAFLR	REPAAEVESVGPE
5. ORUFI03G38130.1	MRPLAIRREARLEPP				SOARDERPEVS				REPAAEVESVGPE
6. Os03t0786600-01	MRPLAIRREARLEPP				SQARDERPEVS				REPAAEVESVGPE
7. SECCE5Rv1G0356100.1.CDS.1	MLARRPRPRPPKPLPL	SAEVOSMGYV	EVKMRWKKD	ASEDAVEN	AHAPPLPPLVS	APLISPSPT	PVSAVSKIPPS	LETSDRRVASELR	REPAAEVESVGPE
8. TraesCS5B02G427200.1.cds1	MRSLARPPRPPKPLPS		EVKMPWKKD	ASEDAVDVI		ADISDSDT	PVSAVSKIPPS	LETSDRRVASELR	
9. scaffold65.530.cds	MQSLAR - LRALEPLPP					ADICOCDT	PVSAVSKICPS	FIPDRRVTSFIR	REPAAEVESMGPO
10. TVU44703	MOSLAK - SRALESLPS				THSRDIRPLAS				REPAAEVESVGPO
11. PUZ36689	MOSLAK - PRALEPLPP				SHARDLRPLAS				REPAAEVESVGPO
12. genblast Os03t0786600-01 Saccharum spon					SHARDIRPLVA				REPAAEVESVG-O
									REPAAEVESVGPE
13. KQK12852				ASEDAVPVL	CMARDLRPLVS		PUSAUSKLERS		REPAAEVESVGPE
14. TKV90959 15. EER90771									
				SSEDAVPEL	SHARDIRPLVS		PVSAVSNLGPL		
16. genblast_Os03t0786600-01_Zea_mays_1	MQSLAK - SRSLESVLP								REPAAEVESMG-Q
4 100000770		20 130 REERDVEAARR		140			170 180		200
1. KQK86779		REERDVEAARR			LPLRVAQGMLW	HLGIPEDYEK	DPDHGTEHDGE		DNDGDGRELGLID
2. HORVU.MOREX.r3.5HG0516920.1.CDS1	HNHPWERLSPSAARLQ							RILTT-GDVVCP-	
3. LPERR03G30940.1	HNHPWERLSDPAARLL							RILTE-GDSVCR-	DEEDDGKELLLID
4. cds.KYUSt_chr4.7551	HHHPWERLSAPAARLL	QEERDVEAARR	ADIASRLRR	LLLMSPARR		HLGLPEDYFR	RPEEDIGQDGEI		KDENDGKELGLID
5. ORUFI03G38130.1	HNHPWFRLSGSAAGLL							RILTI-GDSVCR-	
6. Os03t0786600-01		QEEREVEAARR	ADIISRLGR		LPLRAAQGMLW	HLGLPEDYFR	CRDYDIAQDGEI	RILTI - G <mark>DSVCRE</mark>	
7. SECCE5Rv1G0356100.1.CDS.1		REERDVFAACR		RELLEMSPARR	LPLSVAQGMLW	HLGLPEDYFK	RPDFD G QD G F	RILTT-G <mark>D</mark> VLWS-	EDENHGRELGLID
8. TraesCS5B02G427200.1.cds1		REERDVEAARR				HLGIPEDYFK			EDENHGKELGLID
9. scaffold65.530.cds		REERDVEAARR		LILMCPRRR		HLGIPEDYEK	SPDHGLAQDGE	RILTS-GGGVI	DDDDDGRELGLID
10. TVU44703		REERDVEAARR					EPEYDIVQDGF	RVIAS-RDGTIC-	
11. PUZ36689					LPLRVAQGMLW		DPDQGLAQDGE	QILTS-GDGVICQ	DDDDDGRDLGLID
genblast_Os03t0786600-01_Saccharum_spon	t HNIPWERLSDAAARLL	REERDVEAARR						R I V I S – G D G – – – –	
13. KQK12852	HNHPWERLSGSAARLL	REERDIEAARR	ADITSRLRR	RELE MC PTRR		HEGLPEDYFE	CPYFEIRQDGEI	RILTS-GDVVCR-	EDANDGKELALID
14. TKV90959	HNLPWFRVSDAAARLL	REERDVEAARR	ADITGRERR	LVEMCPRRR		HLGIPEDYFK	DPDHGIEHDGE	RILTS-ADGVICQ RIVIS-GDD	DNDGDGRELGLID
15. EER90771	HNIPWERLSDAAARLL	REERDVFAARR	ADVCGRLRR	VVLMCPRRR		HLGIPEDYFK	DLDHDIAQDGFI	RIVIS-GDD	
16. genblast Os03t0786600-01 Zea mays 1	HNIPWERLSDAATRLL	REERAVEAARR	ADVCARLRR	IVLMTPRRR		HEGIPENYLK	DLDYDIAQDGF	RTVIS-GDG	DHERELELID
	210 220	230	240	250	260	270	280	290	300
1. KQK86779	DGKGQEMPLSVLQM	KEGSMADVP	I P L F P S K G L		LEGEQRLPYVS				AERRELCERQHE
HORVU.MOREX.r3.5HG0516920.1.CDS1	DAKDQEMPLSVLQMGA	MRRSGSAEEVP	FPLFPSKGL	RLKRKIGDW	MEGEQKLPYIS	PYEDFSNIHR	GSDVSEKRAVG	VLHELLSLFVTCS	AERRELICERTHE
3. LPERR03G30940.1	NGEDQELPKSVLEMDA		I P L F Q S K G L	RLKRKIEVW	MEGEQKLPYVS	PYEDFSGIDR	CSDVSEKRVVG	VLHELLSLFVTCS	AERRRLLCLRQHL
4. cds.KYUSt chr4.7551	DVQDQEMPLSVLQTNA	IRREGSADEV P	IPLFPSKGL	RLKHKIGDW	LERFOKLPYVS	PYEDFSNIQP	GTDVSQKRVAG	VLHELFSLFVTCS	AERRRLLCLRTHL
5. ORUFI03G38130.1	NGEQQEMPKSVLQMDA	IRREGSMETV P	IPLFQSK GL	RLKQKIEAW		PYEDFSGIDR	DSDVSEKRVVG	VLHELLSLFVTCS	AERRRLLCLRQHL
6. Os03t0786600-01	NGEHQEMPKSVLQMDA	IRREGSMETVP	IPLFQSKGL	RLKOKIEAW		PYEDFSGIYR	DSDVSEKRVVG	VLHELLSLFVTCS	AERRELCEROHL
7. SECCE5Rv1G0356100.1.CDS.1	DGRDQEMPLSVLQMGA	IRREGSPEEV P	FPLFLSKGL	RLKRELRDW	MEGEQKLPYIS	PYEDFSNIHR	GSDVSEKRAVG	VLHELFSLFVTCS	AERRRLLCLRTHL
8. TraesCS5B02G427200.1.cds1	DGKDQEMPLSVLQMGA	IRREGSAEEV P	FPLFPSKGL	RLKRKIGDW	MEGEOKLPYIS	PYEDFSNTHR	GSDVSEKRAVG	VLHELFSLFVTCS	AERRRLLCLRTHL
9. scaffold65.530.cds	DGKVQEMPLSVLQMNA	MRKSGSMAEVP	IPLFQSK GI	RLKQKIKDW	LEGFORLHYVS	PYEDFSHIRR	GSDVSDKRAVG	VLHELLSLFVTCS	AERRRLLCLRQHL
10. TVU44703	DGKSEEMPLSVLQRNA	KKEGSAAEVP	PLFPSKGL	RLKRKIGDW		PYEDSSNINR	NSDTSDKRVVG	VLHELFSLFVTCS	AERRELCEROHL
11. PUZ36689	DGKLQEMPLSVLQMNA	MRKEGSAAEVP	IPLFQSKGL	RLKQKIKDW	LEGEORLHYVS	PYEDFSHIRP	GSDVSEKRAVG	VLHELLSLFVTCS	AERRRLLCLRQHL
12. genblast Os03t0786600-01 Saccharum spon	t DG <mark>KDEEMPLSVLQLN</mark> A	MRKEGSMEEVS	VPL FPSKG	REKOKEKDW	FREORE PYVS	PYFDESHINR	KRAVG		AERRERCEROHL
13. KQK12852	DGKDQEMPLSVLQMDA		IPLFQSK GL	RLKRKIEDW		PYEDFSNIHR	GSDVSEKRVVG	VLHELFSMFVTCS	AERRELCEROHL
14. TKV90959	DGKGQEMPLSVLQM		PLFPSKGL	REKOKIKDW		PYEDESHIER	GSDVSEKRAVG	VLHELLSLEVTCS	AERRELCEROHL
15. EER90771	DGKDEEMPLSVLQLNA	MRKEGSVEEVS	VPLFPSKGL	REKOKIKDW	LERFORUPYVS	PYEDESHINR	GSDVSEKRAVG	VLHELLSLEVTCS	AFRRERCEROHL
16. genblast Os03t0786600-01 Zea mays 1									AFRRERCEROHL
	310 320	330	340	350	360	370	380	390	400 408
1. KQK86779							CDDCCKDXELE		RCEKNE
	GLPOKEHLVEERHPHV	FYLLLKEKTCE	VVLKEAYMA	GGDTAGEH	PMLELRKKYVE		CRRSGKPVELE		
2. HORVU.MOREX.13.5HG0516920.1.CDS1		EYLLLKEKTCE FYLLLKEKSCE		GGHTSEEH	PMLELRKKYVE PMLEVRSKYAR	MEESRELR	RRRSGKPMOLD		NSAATPS
2. HORVU.MOREX.r3.5HG0516920.1.CDS1 3. LPERR03G30940.1	GLSQKEHRVEERHPHI	FYLLLKEKSCF		GGHTSEEH	PMLEVRSKYAR	MEESRELR	RRRSGKPMOLD		NSAATPS
3. LPERR03G30940.1	GLSOKEHRVEERHPH GLPOKEHRVEERHPHV	FYLLLK <mark>E</mark> KSCF FYLLLK <mark>E</mark> KTCF	VVLKEAYMA VVLKEAYMA	GGHTS EEH	PMLEVRSKYAR PMLEVRRKYAG	MEESREI R MEESREN R	RRRSGKPVQLD CRRSGKPFPSK	PEDQE - SEDSKGV HEDLEQIDDSEGA	NSAATPS NSAPVLS
3. LPERR03G30940.1 4. cds.KYUSt_chr4.7551	GLSQKEHRVEERHPH GLPQKEHRVEERHPHV GLPQKEDRGIERHPH	FYLLLK <mark>E</mark> KSCF FYLLLK <mark>E</mark> KTCF FYLLLK <mark>E</mark> KTCF	VVLKEAYMA VVLKEAYMA VVLKEAYMA	GGHTS EEH GGDTA QEH GGDTA EEH	PMLEVRSKYAR PMLEVRRKYAG PMLAVRSKYAG	MEESREI R MEESREI R MEESREI K	RRRSGKPVQLD CRRSGKPEPSKI RRRSGKPVQLD	PEDQE – SEDSKGM HEDLEQLDDSEGA PEDQEESEDWKDA	NSAATPS NSAPVLS NSIIGTHYF
3. LPERR03G30940.1 4. cds.KYUSt_chr4.7551 5. ORUFI03G38130.1	GL SOKEHRVFERHPHI GL POKEHRVFERHPHV GL POKEDRGIERHPHV GL POKEDRGIERHPHV	FYLLLK <mark>E</mark> K <mark>SC</mark> F FYLLLK <mark>EKTC</mark> F FYLLLK <mark>EKTC</mark> F FYLLLK <mark>E</mark> KTCF	VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYLA	GGHTSIEEH GGDTAIQEH GGDTAIEEH	PMLEVRSKYAR PMLEVRRKYAG PMLAVRSKYAG PMLVVRRKYAG	MEESREI R MEESREI K MEESREI K	RRRSGKPVQLD CRRSGKPFPSK RRRSGKPVQLD CRRSGKPFPSK	PEDQE - SEDSKG HEDLEQIDDSEGA PEDQEESEDWKDA HKDHEQIEDSEGA	NSAATPS NSAPVES NSEGTHYF NSAPEES
3. LPERR03G30940.1 4. cds.KYUSt_chr4.7551 5. ORUF103G38130.1 6. Os03t0786600-01	G SOKEHRVEERHPH G POKEHRVEERHPHV G POKEDRGIERHPHV G POKEHRVEERHPHV G POKEHRVEERHPHV	FYLLLKEKSCE FYLLLKEKTCE FYLLLKEKTCE FYLLLKEKTCE FYLLLKEKTCE	VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYLA VVLKEAYLA	AGGHTS EEH AGGDTA QEH AGGDTA EEH RGDTA EEH RGDTA EEH	PMLEVRSKYAR PMLEVRRKYAG PMLAVRSKYAG PMLVVRRKYAG	MEESROUR MEESROUR MEESROUK MEESROUK MEESROUK	RRRSGKPVQLD CRRSGKPEPSK RRRSGKPVQLD CRRSGKPEPSK CRRSGKPEPSK	PEDQE - SEDSKG HEDLEQIDDSEGA PEDQEESEDWKDA HKDHEQIEDSEGA HKDHEQIEDSEGA	NSAPTS NSAPTS NSAPTS NSAPTS NSAPTS
3. LPERR03G30940.1 4. cds.KYUSL Chr4.7551 5. ORUF103G38130.1 6. OS03t0786600-01 7. SECCE5Rv150356100.1.CDS.1	G SOK HRVFERHPH G POKERRVFERHPH G POKERRVFERHPH G POKERRVFERHPH G SOKERVFERHPH G SOKERVFERHPH	FYLLLKEKSCF FYLLLKEKTCF FYLLLKEKTCF FYLLLKEKTCF FYLLLKEKTCF	VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYLA VVLKEAYLA	AGGHTS EEH AGGDTALQEH AGGDTALEEH RGDTALEEH RGDTALEEH AGGHTSLEEH	PMLEVRSKYAR PMLEVRRKYAG PMLAVRSKYAG PMLVVRRKYAG PMLVVRKYAG	LMESREIR MESREIR MESREIR MESREIR MESREIR MESREIR	RRRSGKPWQLD CRRSGKPEPSK RRSGKPWQLD CRRSGKPEPSK CRRSGKPEPSK RRRSGKPVQLD	PEDQE - SEDSKG HEDLEQIDDSEGA PEDQEESEDWKDA KDHEQIEDSEGA KDHEQIEDSEGA PEDQE - SEDSKGW	NSAATPS NSAPVIS NSIGTHYP NSAPIIS NSAPIIS NSAAIIS
3. LPERR03G30940.1 4. cds.KYUSt_chr4.7551 5. ORUFI03G38130.1 6. OS030786600-01 7. SECCE5RV1G0356100.1.CDS.1 8. TraesCS5B02G427200.1.cds1	GLSOKEHRVERHPH GLPOKEHRVERHPHV GLPOKEDRGIERHPHV GLPOKEHRVERHPHV GLPOKEHRVERHPHV GLSOKEHKVERHPHV GLSOKEHKVERHPHV	FYLLLKEKSCE FYLLLKEKTCE FYLLLKEKTCE FYLLLKEKTCE FYLLLKEKTCE	VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA	AGGHTS EEH AGGDTA QEH AGGDTA EEH RGDTA EEH RGDTA EEH AGGHTS EEH	PMLEVRSKYAR PMLEVRRKYAG PMLAVRSKYAG PMLVVRRKYAG PMLVVRRKYAR PMLEVRSKYAR		RRRSGKPVQLD CRRSGKPFPSK RRRSGKPFPSK CRRSGKPFPSK CRRSGKPVQLD RRRSGKPVQLD	PEDQE - SEDSKG HEDLEQIDDSFGA HKDHEQIEDSFGA HKDHEQIEDSFGA PEDQE - SEDKG	NSAATPS NSAPVIS NSAPVIS NSAPILS NSAATS NSAATS NSAATS
3. LPERR03G30940.1 4. cds, KYUSL chr4.7551 5. ORUFI03G38130.1 6. Os0310786600-01 7. SECCESRV1G0356100.1.CDS.1 8. TraesCS5B02G427200.1.cds1 9. scaff01d65.530.cds	G SOK EHRVE ERHPHI G POK EHRVE ERHPHV G POK EHRVE ERHPHV G POK EHRVE ERHPHV G SOK EHRVE ERHPHV G SOK EHRVE ERHPHV G SOK EHLVE ERHPHV G POK EHLVE ERHPHV	FYLLLKEKSCF FYLLLKEKTCF FYLLLKEKTCF FYLLLKEKTCF FYLLLKEKTCF FYLLLKEKTCF FYLLLKEKTCF	VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA		PMLEVRSKYAR PMLEVRRKYAG PMLAVRSKYAG PMLVVRRKYAG PMLEVRSKYAR PMLEVRSKYAR	MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR	RRRSGKPMQLD CRRSGKPEPSK CRRSGKPEPSK CRRSGKPEPSK CRRSGKPEPSK RRRSGKPMQLD CRRSGKPMQLD CRRGKPTLLE	PEDQE - SEDSKGW HEDLEQIDDSEGA HKDHEQIEDSEGA HKDHEQIEDSEGA PEDQE - SEDLKGW SKVSGPGVIE	NSAATPS NSAATS NSACTHY NSAPILS NSAPILS NSAATS NSAATPS SVEVFS
3. LPERR03G30940.1 4. cds.KYUSt_chr4.7551 5. ORUF003G38130.1 6. 0s03t0786600-01 7. SECCESRv160356100.1.CDS.1 8. TraesCS5B02G427200.1.cds1 9. scaff01d65.530.cds 10. TVU44703	G SQK EHRVE RHPHU G PQK EHRVE RHPHU G PQK EHRVE RHPHU G PQK EHRVE RHPHU G SQK EHRVE RHPHU G SQK EHRVE RHPHU G SQK EHRVE RHPHU G SQK EHLVE RHPHU G SQK EHLVE RHPHU G PQK EHLVE RHPHU	FYLLLKEKSCF FYLLKEKTCF FYLLKEKTCF FYLLKEKTCF FYLLKEKTCF FYLLKEKTCF FYLLKEKTCF FYLLKEKTCF	VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA		PMLEVRSKYAG PMLAVRSKYAG PMLAVRSKYAG PMLVVRRKYAG PMLVVRKYAG PMLEVRSKYAR PMLEVRSKYAR	MEESREIR MEESREIR MEESREIR MEESREIR MEESREIR MEESREIR MEESREIR MEESREIR	RRSGKPVQLD CRRSGKPVQLD CRRSGKPVQLD CRRSGKPPSK CRRSGKPVQLD RRRSGKPVQLD CRRSGKPVQLD CRRSGKPVQLD CRRSGKPVELS	PEDQE - SEDSKG HEDLEQIDDSGA HEDLEQIDDSGA HKDHEQIEDSGA HKDHEQIEDSGA PEDQE - SEDKGW SKVSGPGUE	NSAATPS NSAPVES NSIGTHYF NSAPIES NSAPIES NSAAIPS SAAIPS SAAIPS SAAIPS SAAIPS
3. LPERR03G30940.1 4. cds.KVUSt_Cht4.7551 5. ORUFI03G38130.1 6. Os030786600-01 7. SECCESRv1G0356100.1.CDS.1 8. TraesCS5B02G427200.1.cds1 9. scaff01d65.530.cds 10. TVU44703 11. PU236689	G SOK EHRVE ERHPHI G POK EHRVE ERHPHV G POK EHRVE ERHPHV G POK EHRVE ERHPHV G SOK EHRVE ERHPHV G SOK EHRVE FRHPHV G SOK EHRVE FRHPHV G POK EH VE ERHPHV G POK EH VE ERHPHV G POK EH VE ERHPHV	FYLLEK EKSCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF	VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYLA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA		PML EVR SKYAR PML AVR SKYAG PML AVR SKYAG PML VVR KKYAG PML EVR SKYAR PML EVR SKYAR PML EVR KKYVE PML EVR KKYVE	MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR	RRRSGKPMQLD CRRSGKPMQLD CRRSGKPMQLD CRRSGKPMPPSK RRRSGKPMQLD RRRSGKPMQLD CRRGGKPMQLD CRRSGKPMQLD CRRSGKPMQLD CRRSGKPMQLD CRRSGKPMQLD	PEDQE SEGA PEDQE SEQA PEQQE SEQA PEQQE SEQA PEQQE SEQA PEQQE SEQA PEQQE SEQA PEQQE SEQA SKVSGPGM SKG SNVSGSSGS SCG	NSAATPS NSAPVES NSIGTHYF NSAPIES NSAPIES NSAAIPS SAAIPS SAAIPS SAAIPS SAAIPS
3. LPERR03G30940.1 4. cds.KYU5t_chr4.7551 5. ORUF103G38130.1 6. S003t0786600-01 7. SECCESRv1G0356100.1.CDS.1 8. TraesCS5B02G427200.1.cds1 9. scaff0d65.530.cds 10. TVU44703 11. PUZ36689 12. genblast_cS033t0786600-01_Saccharum_spon	G SOK EHRVE EHHPH G POK EHRVE EHHPH G SOK EHV EHHPH G POK EHV EHHPH G POK EHVE EHPHT G POK EHVE EHHPHT G POK EHVE EHHPHT	FYLLEK EKSCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF FYLLEK EKTCF	VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA	GGGTALEE GGCTALEE RGCTALEE RGCTALEE RGCTALEE GGTALEE GGCTALEE RGCTALEE RGCTALEE RGCTALEE	PML EVR SKYAR PML AVR SKYAG PML AVR SKYAG PML VVR KYAG PML EVR SKYAR PML EVR SKYAR PML EVR KKYVE PML EVR KKYVE PML EVR KKYVE	MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MEESRELR MESRELR MESRELR	RR S GK P MOLD CRR S GK P MOLD CRR S GK P MOLD CRR S GK P P S K CRR S GK P P S K CRR S GK P MOLD CRR S GK P MOLD CRR S GK P MOLD CRR S GK P MOLD CRR S GK P MOLD	PEDQE - SEDSKGM HEDDLEQIDSEGA PEDQEESCLWKDA HKDHUQIEDSEGA HKDHUQIESKGKG SKVSGRGWHE SKGSGCHHE SKSSGSGNHE	NSAPUS NSAPUS NSAPUS NSAPUS NSAPUS NSAAUS NSAAUS NSAAUS SWEVES NSAAUS SWEVES NSAUS SCIES
3. LPERR03G30940.1 4. cds.KVUSt.chr4.7551 5. ORUFI03G38130.1 6. Ob310786600-01 7. SECCE5Rv1G0356100.1,CDS.1 8. TraesCS5802G427200.1,cds1 9. scaff0d65.530.cds 10. TVU44703 11. PUZ36689 12. genblast_0S0310786600-01_Saccharum_spon 13. KQR12852	G SOK EHRVE ERHPHI G POK EHRVE ERHPHY G POK EHRVE ERHPHY G POK EHRVE ERHPHY G SOK EHRVE ERHPHY G SOK EHRVE ERHPHY G SOK EHRVE ERHPHY G POK EH VE ERHPHY		VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA		PMLEVRSKYAR PMLEVRSKYAG PMLAVRSKYAG PMLVRKYAG PMLEVRSKYAR PMLEVRSKYAR PMLEVRSKYAR PMLEVRSKYZ PMLEVRSKYZ PMLEVRSKYZ PMLEVRSKYZ	MESRENK MSRENK MSRENK MSRENK MSRENK MSRENK MSRENK MSRENK MSRENK MSRENK	RRRSGKPMQLD CRRSGKPHPSK CRRSGKPHPSK CRRSGKPHPSK CRRSGKPHPSK CRRSGKPHPSK CRRSGKPHPSK CRRSGKPHPSK CRRSGKPHPSK CRRSGKPHP CRRSGKPHPF CRRSGKPHPF	PEDQQ - SEDSKGW PEDQQUSSEDWKDA PEDQQUSSEDWKDA PEDQQUSSEDWKDA PEDQQ - SEDWKGW PEDQQ - SEDWKGW PEDQQ - SEDWKGW SEDSKGW SINJSGSGUL SINJSGSGUL SINJSGSGUL PUDHWESEDSEGW	INSAATPS NSAPUS NSAPUS NSAPUS NSAPUS NSAATUS NSAATUS NSAATUS NSAMES DCIETUS NSASUFURS
3. LPERR03G30940.1 4. cds, KYU35, Chr4.7551 5. ORUF103G38130.1 6. Os030786600-01 7. SECCESRv1G0356100.1.CDS.1 8. TraesCS5B02G427200.1.cds1 9. scaff10d65.530.cds 10. TVU44703 11. PUZ36689 12. genblast_Os0310786600-01_Saccharum_spon 13. KQK12852 4. TKY90559	G SOK EHRVE EHPHU G POK EHRVE EHPHU G POK EHRVE EHPHU G POK EHRVE EHPHU G SOK EHRVE EHPHU G SOK EHKVE EHPHU G SOK EHKVE EHPHU G POK EH VE EHPHU G POK EHVE EHPHU G POK EHVE EHPHU		VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA VVLKEAYMA		PML EVR SKYAR PML ANR SKYAG PML ANR SKYAG PML VYR KYAG PML EVR SKYAR PML EVR SKYAR PML EVR KKYVE PML EVR KKYVE PML EVR KKYVE PML EVR SKYAR		RRRSGKPMQLD CRRSGKPMQLD CRRSGKPMQLD CRRSGKPMQLD RRRSGKPMQLD CRRSGKPMQLD CRRSGKPMQLD CRRGKPTMLD CRRSGKPMQLD CRRSGKPMQLD CRRSGKPMQLD CRRSGKPMQLD CRRSGKPMQLD CRRSGKPMQLD CRRSGKPMQLD	P DOQ - S DD S K GM P DOQ - S DD S F GA P DOQ - S DOD S F GA H K D H O D S S DOL K GM H K D H O D S S GA H K D H O D S S GA S M S C P C D S K GM S K M S C P C D S K GM S K M S C S C D L K GM S C S C S C S C D L K GM S C S C S C S C D L K GM S C S C S C S C D L K GM S C S C S C S C S C S C S C S C S C S C	INSAATPS NSAPUS NSAPUS NSAPUS NSAPUS NSAATUS NSAATUS NSAATUS NSAMES DCIETUS NSASUFURS
3. LPERR03G30940.1 4. cds.KVUSt_chr4.7551 5. ORUFI03G38130.1 6. Op30786600-01 7. SECCE5RV160356100.1.CDS.1 8. TraesCS5B02G427200.1.cds1 9. scaffold65.530.cds 10. TVU44703 11. PU236689 12. genblast_OS030786600-01_Saccharum_spon 13. KQK12852 14. TKV90959 15. EER90771	G POKENKERVERHPHU G POKENKGIRHPHU G POKENKGIRHPHU G POKENKIRHPHU G POKENKIRHPHU G POKENKIRHPHU G SOKENKIRHPHU G SOKENLVERHPHU G POKENLVERHPHU G POKEN VERHPHU G POKEN VERHPHU G POKENLVERHPHU G POKENLVERHPHU G POKENLVERHPHU				PMLEVRSKYAR PMLEVRKYAG PMLEVRKYAG PMLEVRKYAG PMLEVRKYAG PMLEVRKKYVE PMLEVRKKYVE PMLEVRKKYVE PMLEVRKKYVE PMLEVRKKYVE PMLEVRKKYVE PMLEKKKYVE		RRSGKPWQLO CRSGKPPSK RRSGKPPSK CRSGKPPSK CRSGKPPSK RRSGKPQLO CRGKPTCC CRSGKPTCC CRSGKPLESY CRSGKPLESY CRSGKPLESY CRSGKPLESY CRSGKPLESY CRSGKPLESY	P DOQ - S DO S KGM P DOQ - S DO S KGM P DOQ - S DOWK DA HKD HKQ HAD S G A HKD HKQ HAD S G A P DOQ - S DO S KGM S KV S G P G U H S NV S G S G DO S KGM S KW S G S G DO S KGM S KW S G N N M S DO S G S G DO S KGM S KW S M N T DO P D K - S KW S M N	INSAATPS NSAPUS NSAPUS NSAPUS NSAPUS NSAATUS NSAATUS NSAATUS NSAMES DCIETUS NSASUFURS
3. LPERR03G30940.1 4. cds, KYU35, Chr4.7551 5. ORUF103G38130.1 6. Os030786600-01 7. SECCESRv1G0356100.1.CDS.1 8. TraesCS5B02G427200.1.cds1 9. scaff10d65.530.cds 10. TVU44703 11. PUZ36689 12. genblast_Os03t0786600-01_Saccharum_spon 13. KQK12852 4. TKY90559	G SOK EHRVE EHPHU G POK EHRVE EHPHU G POK EHRVE EHPHU G POK EHRVE EHPHU G SOK EHRVE EHPHU G SOK EHKVE EHPHU G SOK EHKVE EHPHU G POK EH VE EHPHU G POK EHVE EHPHU G POK EHVE EHPHU				PMLEVRSKYAR PMLEVRKYAG PMLEVRKYAG PMLEVRKYAG PMLEVRKYAG PMLEVRKKYVE PMLEVRKKYVE PMLEVRKKYVE PMLEVRKKYVE PMLEVRKKYVE PMLEVRKKYVE PMLEKKKYVE		RRSGKPWQLO CRSGKPPSK RRSGKPPSK CRSGKPPSK CRSGKPPSK RRSGKPQLO CRGKPTCC CRSGKPTCC CRSGKPLESY CRSGKPLESY CRSGKPLESY CRSGKPLESY CRSGKPLESY CRSGKPLESY	P DOQ - S DO S KGM P DOQ - S DO S KGM P DOQ - S DOWK DA HKD HKQ HAD S G A HKD HKQ HAD S G A P DOQ - S DO S KGM S KV S G P G U H S NV S G S G DO S KGM S KW S G S G DO S KGM S KW S G N N M S DO S G S G DO S KGM S KW S M N T DO P D K - S KW S M N	INSAATPS NSAPUS NSAPUS NSAPUS NSAPUS NSAATUS NSAATUS NSAATUS NSAMES DCIETUS NSASUFURS

Group: Os02t0763000-01 max_score: 402

	1 10	20	30	40	50	60	70	80	90	100
1. KQK01766	MYPAAPADAY-DRYSN	- GTPPPSAPAP-	APY-	<mark>OH</mark> A	MHQGRPAC	GGLTRWSTG	FHCMDDPGNC			
2. HORVU.MOREX.r3.6HG0617750.1		- GPPPTAPPPP-		- – – <mark>NH</mark> – – – A	MNOSHHAHPAA	AGLARWSTG	FHCMDDPGNO			
3. LPERR02G27500.1	MYPSAPPDAY - NTYSA		ATY-	<mark>00</mark> K	MNTPRPG	GETRWSTG	EHCMDDPGNO			
4. genblast_Os02t0763000-01_Lolium_perenne_6		NGTPMPAAPPO-	ATY-	NH A	MN HSRPAA	AGERRWSTS	FHCMDDPGNO			
5. ORUFI02G35140.1	MYPSAPPDAY-NKYSA				MN TPRTGO		EHCMDDPGNC			
6. Os02t0763000-01	MYPSAPPDAY - NKYSA			O P T	MN TPRTGO	GETRWSTG	EHCMDD PGNC			
7. SECCE6Rv1G0410970.1	MYPTTPSDAY-DRESS	-GPPPTAPPPQC			ON-HHHGHPAA					
8. TraesCS6A02G321000.1	MYPTTPSDAY-DRESS	-GPPPTAPPPO-			MNOSHHARPA					
9. scaffold55.23.cds	MYPSAPPDAY-NKESS	-GAPPTAPPPP-			MN SSRPGO		FUCMDDRCNC			
10. TVU28621	MYPSAPPDAY-NKEST	-GAPPTAPPP			MNPARTGA	AGTKWSTG	EHCMDDPGNC	KSFYFSSSSSSQ		
11. PUZ77764	MYPSAPPDAY-NKENT	-GAPPTAPPP			MN PSRPGA					
12. Sspon.04G0022300-2C-mRNA-1:cds	MYPSAPPDAY-NKESS				MN PSRPGO		FHCMDDPGNC			
13. KQL31632	MYPSAPPDAY-NKETT	-GAPPTAPPP			MN P SRPG		EHCMDDPGNC			
14. TKW41734	MYPSAPPDAY-NKETT	-GAPPTAPPP			MN P SRPG		FHCMDDPGNC			
15. EES05849	MYPSAPPDAYSNKESS				MN PSRPGO					
16. Zm00001eb193700_P001	MYPSAPPDAY-NKESA					GERKWSTG				
10. ZIII00001ED193700_F001		120 130		140	150	160	170	180	190	200
1. KOK01766										
2. HORVU.MOREX.r3.6HG0617750.1		PCITE COLORIN		CAGSGAA			MRAHYDLDE	ECPDELVHWCCE		KNRGEDMG
3. LPERR02G27500.1		PCTTEGOVADIN		CLASGSVY						
4. genblast_Os02t0763000-01_Lolium_perenne_6								FCPDFLVHWCCF		
5. ORUFI02G35140.1					ALLCAS-GMG					
6. Os02t0763000-01					ALLCAS-GMG					
7. SECCE6Rv1G0410970.1								ECPDFLVHWCCE		KNRGFDMG
8. TraesCS6A02G321000.1										
9. scaffold55.23.cds		PCITEGOTADIN								
					G CASTGMG					
10. TVU28621 11. PUZ77764			DICTC		GLICASTGMG					
					G CASTGMG					
12. Sspon.04G0022300-2C-mRNA-1:cds										
13. KQL31632										
14. TKW41734									HLALCQEYREL	
15. EES05849				PCLASGETY						
16. Zm00001eb193700_P001				CLASGL VY	GLICASTGMG					
1 KOK01766	210 220	230	239							
1. KQK01766	GWDANMERORRGVA-									
2. HORVU.MOREX.r3.6HG0617750.1			GIVINI							
3. LPERR02G27500.1		GAQVMGA-PALP	VGIVIVIR							
4. genblast_Os02t0763000-01_Lolium_perenne_6										
5. ORUFI02G35140.1	GWAANVDRORRGVT -									
6. Os02t0763000-01	GWAANVDRORRGVT-									
7. SECCE6Rv1G0410970.1	GWDANMERRNRGVT -									
8. TraesCS6A02G321000.1	GWDANMERRNR GVT -									
9. scaffold55.23.cds	GWDANMDRORRGVA-									
10. TVU28621	GWDANMERORRGVS -		HGMMR							
11. PUZ77764	GWDANMDRORRGVA-		LGMIR							
12. Sspon.04G0022300-2C-mRNA-1:cds		GGTVMGA - PALF								
13. KQL31632	GWDANVDRORRGVA-									
14. TKW41734	GWDANVDRORRGVA-									
15. EES05849										
16. Zm00001eb193700_P001	GWEANMDRORRGVAC	GGAMIGAPPALE								

Figure 3.JPEG

