dbWFA: a web-based database for functional annotation of Triticum aestivum transcripts

Jonathan J. Vincent, Zhanwu Z. Dai, Catherine Ravel, Frédéric Choulet, Saïd S. Mouzeyar, Mohamed-Fouad M. Bouzidi, Marie M. Agier, Pierre Martre

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

Database tool

dbWFA: a web-based database for functional annotation of *Triticum aestivum* transcripts

Jonathan Vincent[1,2,3], Zhanwu Dai[1,2], Catherine Ravel[1,2], Frédéric Choulet[1,2], Said Mouzeyar[1,2], M. Fouad Bouzidi[1,2], Marie Agier[3] and Pierre Martre[1,2,*]

[1]INRA, UMR1095 Genetics, Diversity and Ecophysiology of Cereals, 5 Chemin de Beaulieu, Clermont-Ferrand, F-63 039 Cedex 2, France, [2]Blaise Pascal University, UMR1095 Genetics, Diversity and Ecophysiology of Cereals, Aubière F-63 177, France and [3]Blaise Pascal University, UMR6158 CNRS LIMOS Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, Aubière F-63 173, France

Present address: Zhanwu Dai, INRA, ISVV, UMR1287 Écophysiologie et Génomique Fonctionnelle de la Vigne (EGFV), F-33 882 Villenave d'Ornon, France

***Corresponding author:** Tel: +33 473 624 351; Fax: +33 473 624 457; Email: pierre.martre@clermont.inra.fr

Submitted 11 July 2012; Revised 19 February 2013; Accepted 20 February 2013

Citation details: Vincent,J., Dai,Z.W., Ravel,C. *et al*. dbWFA: a web-based database for functional annotation of *Triticum aestivum* transcripts. *Database* (2013) Vol. 2013: article ID bat014; doi:10.1093/database/bat014

The functional annotation of genes based on sequence homology with genes from model species genomes is time-consuming because it is necessary to mine several unrelated databases. The aim of the present work was to develop a functional annotation database for common wheat *Triticum aestivum* (L.). The database, named dbWFA, is based on the reference NCBI UniGene set, an expressed gene catalogue built by expressed sequence tag clustering, and on full-length coding sequences retrieved from the TriFLDB database. Information from good-quality heterogeneous sources, including annotations for model plant species *Arabidopsis thaliana* (L.) Heynh. and *Oryza sativa* L., was gathered and linked to *T. aestivum* sequences through BLAST-based homology searches. Even though the complexity of the transcriptome cannot yet be fully appreciated, we developed a tool to easily and promptly obtain information from multiple functional annotation systems (Gene Ontology, MapMan bin codes, MIPS Functional Categories, PlantCyc pathway reactions and TAIR gene families). The use of dbWFA is illustrated here with several query examples. We were able to assign a putative function to 45% of the UniGenes and 81% of the full-length coding sequences from TriFLDB. Moreover, comparison of the annotation of the whole *T. aestivum* UniGene set along with curated annotations of the two model species assessed the accuracy of the annotation provided by dbWFA. To further illustrate the use of dbWFA, genes specifically expressed during the early cell division or late storage polymer accumulation phases of *T. aestivum* grain development were identified using a clustering analysis and then annotated using dbWFA. The annotation of these two sets of genes was consistent with previous analyses of *T. aestivum* grain transcriptomes and proteomes.

Database URL: urgi.versailles.inra.fr/dbWFA/

Introduction

Triticum aestivum (L.), common wheat or bread wheat, is one of the most important staple crops in the world. It is cultivated worldwide and provides >20% of the calories and proteins in the human diet (http://faostat.fao.org). Although ongoing sequencing efforts have already produced important genomic resources (1–4), the complete sequencing and annotation of the hexaploid (2n = 6× = 42, AABBDD) *T. aestivum* genome has yet to be achieved. A first version of the genome of the bread wheat cv. Chinese Spring has recently been published (4), providing the scientific community with highly valuable genomic and evolutionary information, which will facilitate

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

genome-wide analysis of bread wheat. However, because of the low-coverage (5-fold) shotgun sequencing method used in this project, this resource does not represent a high-quality draft of the wheat genome in terms of sequence completion, quality and annotation. Genome-wide analysis of gene expression by means of expression microarrays or transcriptome sequencing (RNA-Seq) is now being adopted in *T. aestivum* (5, 6), but analysing such large datasets requires extensive annotation efforts. Data fragmentation and technical and semantic heterogeneity can severely limit the efficient extraction and interpretation of biological data (7, 8).

More and more genomic information is becoming available for *T. aestivum* research. Various resources and associated tools grant the user a structural overview of expressed sequence tag (EST) (ITEC, http://avena.pw.usda.gov/genome/) (9, 10) or bacterial artificial chromosome clone libraries (11, 12) for instance (13–15). Important initiatives are underway to facilitate the breeding of improved Triticeae varieties. The TriticeaeGenome project (www.triticeaegenome.eu) grants access to comprehensive information extracted from experimental data to provide a better understanding of Triticeae genomes (16). The global database GrainGenes (http://wheat.pw.usda.gov/) provides a variety of services and bioinformatics tools for the Triticeae and *Avena sativa* research communities. The HarvEST database (http://harvest.ucr.edu/) (17), dedicated to several crop species, including *T. aestivum* and *Hordeum vulgare*, provides access to curated EST assemblies, comparative analysis tools and links to orthologues in related model plant species. Together, these resources compile and cross-reference a great deal of information on physical and genetic mapping, markers, sequence variations and quantitative trait loci. To some extent, they also provide information leading indirectly to predicted gene product functions, but none of them is focused on functional gene annotation, and it is necessary to navigate through numerous unlinked resources to extract functional information.

Recently, pipelines for the automated annotation of genomic sequences of *T. aestivum* and related species have been developed (3, 18). These pipelines are based on the prediction of gene models within genome sequences, so they are not able to functionally annotate sequences originating from transcriptome sequencing like ESTs. Because no reference genome sequence is yet available for *T. aestivum*, a massive sequencing effort has produced more than a million ESTs (http://wheat.pw.usda.gov/genome/). To deal with the high level of redundancy of this resource, these sequences were clustered (i.e. overlapping and partial polyA-tailed expressed sequences are grouped) to provide a reference set of unique expressed genes, NCBI UniGenes (http://www.ncbi.nlm.nih.gov/UniGene). The assembly conditions used to build NCBI UniGenes make it the most comprehensive coding DNA (cDNA) assembly available to date, and UniGene assemblies have been used as a reference set of sequences for many species. An additional effort was made to construct full-length cDNA sequences that were included in the TriFLDB database (19). Full-length cDNA sequences are most commonly used in genome annotation as a resource for cross-species comparative analyses. Currently, TriFLDB is the most reliable source of full-length cDNA sequences in *T. aestivum*. TriFLDB includes annotations based on homologies found by searching protein databases, extensive Gene Ontology (GO) annotations and InterProScan results. Recently, a new collection of nearly 1 million ESTs, assembled into contigs and singlets, was annotated with GO terms (20), but meaningful prediction of gene function requires more than one system of annotation.

After the sequencing of the first plant genome, *Arabidopsis thaliana* in 2000 (21), several plant sequencing projects have been successful. The sequenced genomes most closely related to the *T. aestivum* genome are those of *Oryza sativa* ssp. *indica* (22), *Oryza sativa* ssp. *japonica* (23), *Zea mays* (24), *Glycine max* (25), *Sorghum bicolor* (26), *Brachypodium distachyon* (27) and *Hordeum vulgare* (28). Both structural and functional annotation resources for these species are developing steadily. One of the most effective methods to annotate a transcript is to find its orthologous counterparts in well-annotated closely related genomes (29, 30). Although *H. vulgare* (L.) would be expected to be the most useful reference because it is more closely related to *T. aestivum*, comprehensive and high-quality annotations of gene function are only available for *O. sativa* and *A. thaliana* (2), essentially owing to the lengthy and accurate annotation efforts undertaken.

To annotate ESTs or transcripts using sequence homology, it is necessary to navigate through unrelated databases. Some tools using this homology approach have been developed. For instance, Blast2GO (31) can be queried using *T. aestivum* sequences to give GO results. ONDEX (7), developed with the challenges of functional annotation of *T. aestivum* genome in mind, combines data integration from various sources and various mining methods, including graph-based analyses, to annotate wheat gene functions according to a wisely chosen set of annotation standards. However, ONDEX does not provide easy access to workable static results that are often required in research. To fill this gap, we developed dbWFA, an open-access database relating the *T. aestivum* UniGene set and the full-length cDNA sequences from TriFLDB to *A. thaliana* (TAIR10) (32) and *O. sativa* (pseudomolecules version 7.0) (33) annotation through BLAST (34) results. dbWFA also includes the inventory database of *T. aestivum* transcription factors (wDBTF) (35) and hand-curated gene families (36). As an all-in-one interface for the annotation of *T. aestivum* sequences, dbWFA will be

useful to the researchers working on *T. aestivum* and more generally on cereals, particularly for comparative cereal genomics and functional genomics. The web implementation of dbWFA provides an easy-to-use interface to annotate transcript sequences from *T. aestivum*, with functional information from multiple pervasive annotation systems. Here, the use of dbWFA is illustrated with several query examples, and the quality of the annotation method is assessed by comparing the MapMan bin annotation of all the transcripts of the *T. aestivum* NimbleGen 40 k microarray (37) with that of *A. thaliana* and *O. sativa*. The use of dbWFA is further illustrated by analysing the annotation of 433 genes specifically expressed during either the early cell division or the late storage polymer accumulation (SPA) phases of grain development.

Data Content, Database Architecture and Web Interface

Five functional classification/annotation systems were integrated (Figure 1) in dbWFA to offer a fast and efficient functional annotation tool for *T. aestivum* UniGenes:

- GO (http://www.geneontology.org) (38), a non-redundant structured hierarchy of ontologies, which is the most widely used functional annotation system in bioinformatics. The GO project provides an efficient annotation standard that can be applied to numerous species. It is built on a controlled vocabulary of terms for describing gene function. dbWFA includes GO annotation data (OBO version 1.2) for *A. thaliana* and *O. sativa.*

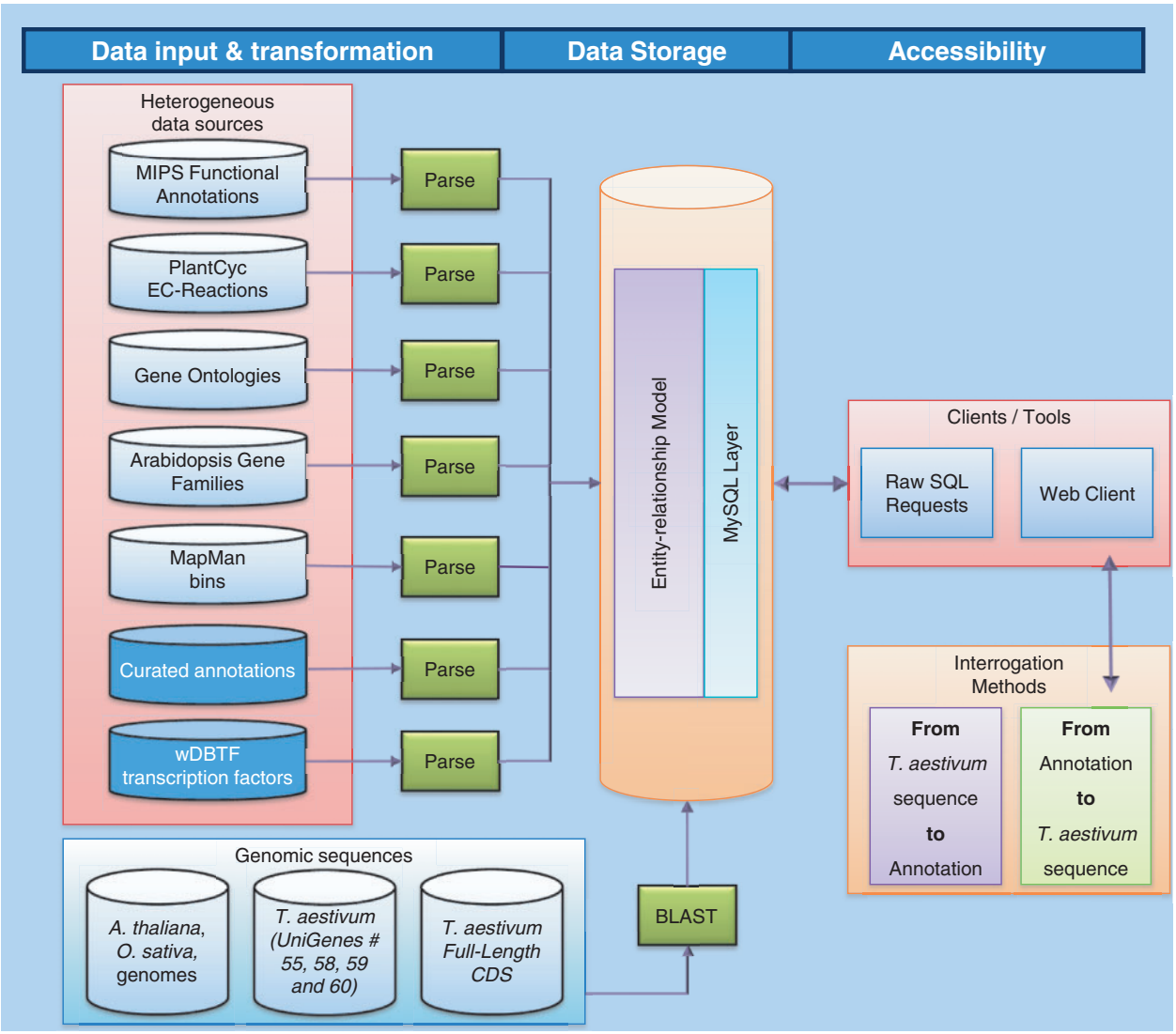

Figure 1. Simplified diagram of the data integration process.

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

- Plant Metabolic Network (PMN; http://www.plantcyc.org) (39), which provides a broad network of curated databases on primary and secondary plant metabolism, including pathways, enzymes, genes, compounds and reactions from several plant species. dbWFA contains data from AraCyc (version 9.0) for *A. thaliana* and RiceCyc (version 3.2) for *O. sativa*.
- MapMan (http://mapman.gabipd.org) (40), which is a user-driven tool for large datasets (e.g. gene expression data from microarrays) visualized in the context of diagrams of metabolic pathways or other processes. MapMan annotation data (bin tree version 1.1) for both *A. thaliana* and *O. sativa* are stored in dbWFA. The dbWFA database also provides a function to automatically generate MapMan *T. aestivum* mapping files.
- Munich Information Center for Protein Sequences Functional Catalogue (MIPS FunCat; http://www.helmholtz-muenchen.de/en/mips/projects/funcat) (41), which provides a hierarchical scheme for the functional description of proteins of prokaryotic and eukaryotic origin. MIPS FunCat annotations for *A. thaliana* (MAtDB version 2.1) are stored in dbWFA.
- *A. thaliana* Gene Family Information (TAIR version 10; http://www.arabidopsis.org/browse/genefamily) (42), which provides gene family information for the plant model species *A. thaliana*.

The 17 541 full-length cDNA sequences from the TriFLDB and other public databases and all the transcript sequences from the *T. aestivum* UniGene set (builds #55, #58, #59 and #60) were processed using the BLASTx algorithm against *A. thaliana* and *O. sativa* predicted cDNA sequences (Figure 1). Build #55 (the one used to develop the *T. aestivum* NimbleGen 40 k microarray) (37) and the following major releases were retained, as users may have developed resources based on different builds of the UniGene even though NCBI only stores the most recent build. BLAST results with an e-value $>10^{-3}$ were not stored in the database, as we considered this would be too poor a match for most research. No other filter was applied to the BLAST results before their insertion into the database. All the parameters from the BLAST tabular results were kept, and $>30 \times 10^6$ BLAST results for the UniGenes and 95×10^6 BLAST results for the ESTs shaping the UniGene clusters were stored, so they could be rapidly screened when querying the database.

The database also contains curated information on *T. aestivum* transcription factors (2891 transcripts), E3 ubiquitin ligases of the ubiquitin-proteasome system (876 transcripts), hormone-responsive genes (467 transcripts) and seed storage proteins (55 transcripts; Figure 1). Transcription factor UniGenes were retrieved from the wDBTF database (34). E3 ligase and hormone-responsive UniGenes were recovered from the NCBI and TAIR databases using all *A. thaliana* and *O. sativa* E3 ligase and hormone-responsive sequences as the query in homology searches using the BLASTn, BLASTx and tBLASTx programs (36). The BLAST hits were filtered using an e-value threshold of 10^{-5} and an alignment length exceeding 80 bp. All sequences were checked for consistency and for the presence of specific protein signatures using the InterProScan program (http://www.ebi.ac.uk/Tools/pfa/iprscan/). For seed storage protein UniGenes, homology searches were performed on the whole UniGene build #55, using BLASTx and *T. aestivum* seed storage protein sequences as reference. No preliminary filter was applied to BLASTx results. Instead, all the alignments were carefully examined, and similarity in known conserved critical regions of seed storage proteins was given priority over e-value and BLAST score alone. In dbWFA, curated UniGene annotations are assigned to *T. aestivum* transcripts without any intermediate BLAST result.

Following the recommendations of the International Wheat Genome Sequencing Consortium (IWGSC) for annotating *T. aestivum* genomic sequences (3), the percentages of coverage (with respect to the length of the orthologous proteins) and identity are used to assign functional annotations to a transcript. In dbWFA, users can define the value of these two parameters, but we strongly recommend using the cutoff values suggested by the IWGSC, where BLAST results with an identity >45% and coverage >50% are assigned a 'putative function' and BLAST results with identity and coverage >90% are assigned a 'known function'.

All data are stored in a MySQL database. The integration of the database allows one to assign the functional annotation from any of the systems described above to the transcripts of interest and vice versa. The dbWFA database thus provides a very powerful resource for the annotation of *T. aestivum* UniGenes. To find the most commonly sought types of information from dbWFA, simple yet pertinent queries with their parameters can be sent through a web-based interface (Figure 2). The results are delivered as html pages, and an export procedure is available to retrieve data in spreadsheet. The html result pages provide links redirecting the user to websites of the different annotation systems, allowing a global analysis of the annotation results. The web interface can also be used to automatically create MapMan mapping files for the search results. Although the dbWFA web interface only allows data mining of common queries, specific queries can be performed using the SQL database, which can be downloaded from the dbWFA website. The modularity of the database will facilitate the integration of new *T. aestivum* data as transcripts are sequenced and annotated through different pipelines.

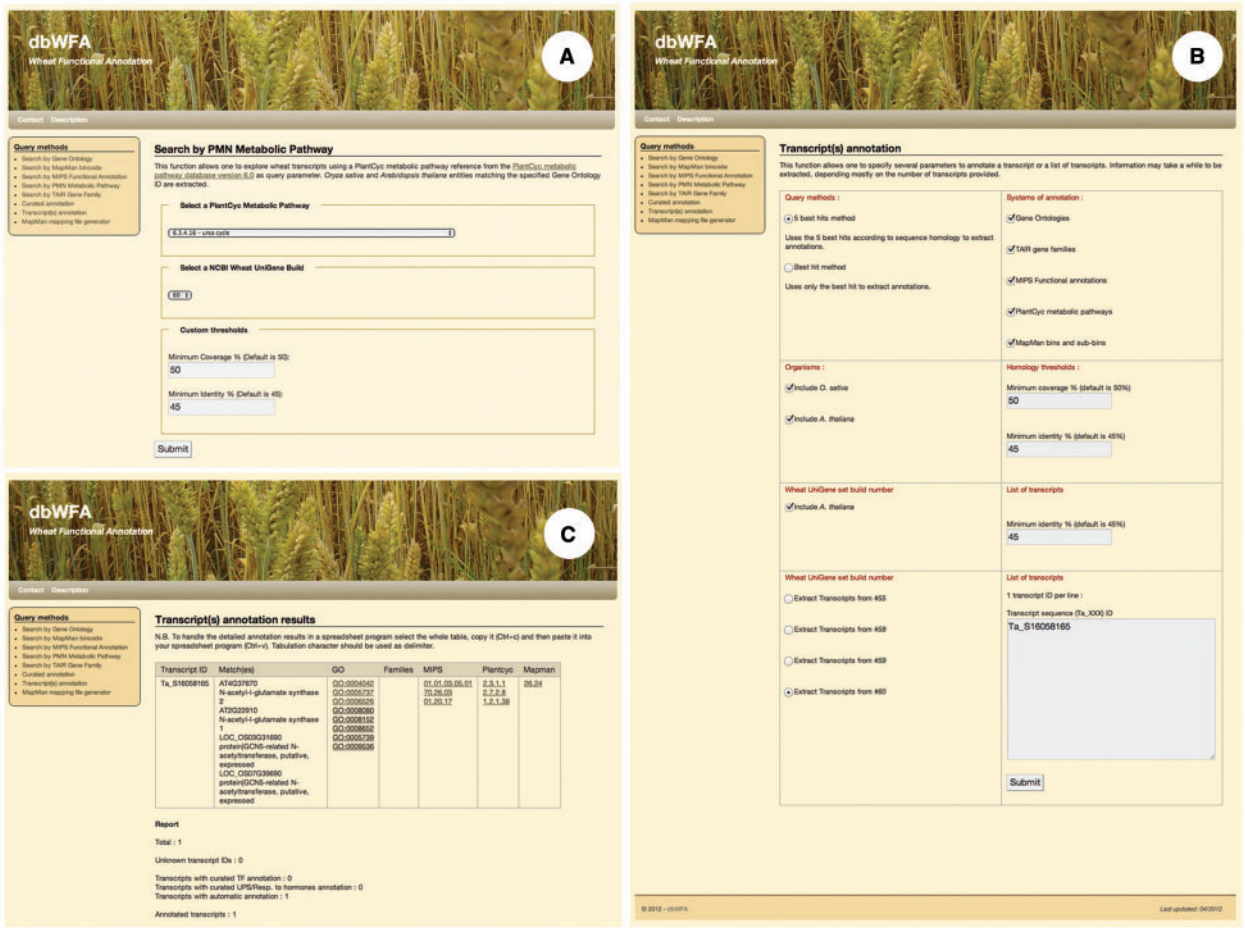

Figure 2. Screen capture of the web interface of the dbWFA database. **(A)** Page for querying PMN pathways. Similar pages can be used to query the MIPS Functional Category, TAIR gene families, GO and MapMan bins. A list of GO can be queried simultaneously. **(B)** Page for querying UniGene or Full-length cDNA sequences annotations. **(C)** Result page for annotated UniGenes.

Using dbWFA: Percentage of Annotated UniGenes, Comparison of *T. aestivum* UniGene and *A. thaliana* and *O. sativa* Whole-Genome Annotation and Query Examples

Thirty-four percent (13 713 transcript sequences), 40% (14 843), 35% (20 016) and 35% (20 034) of the transcript sequences of the UniGene builds #55, #58, #59 and #60, respectively, have a putative functional annotation in at least one of the annotation resources. Eighty-one percent of the 17 541 full-length cDNA sequences from TriFLDB have a putative functional annotation in at least one of the annotation resources. The number of transcripts and full-length cDNA sequences annotated in the different resources are given in Table 1. BLASTn analysis revealed that 12 478 full-length cDNA sequences matched a sequence in the UniGene set (build #60) with a coverage and identity threshold value >50 and 90%, respectively. Among these 12 478 correspondences, 10 996 and 5 932 full-length cDNA sequences and UniGene sequences, respectively, have a putative functional annotation in at least one of the annotation resources. This result highlights the additional information brought by the full-length cDNA sequences.

The quality of the annotation method is illustrated by comparing the MapMan bin annotation of all the transcripts of the *T. aestivum* NimbleGen 40 k microarray (developed with UniGene build #55) and the full-length cDNA sequences from TriFLDB with the annotation of *A. thaliana* and *O. sativa* imported from MapMan and recorded in the database. The MapMan bins were used here because this annotation system is available for the three species. Overall, there was no clear bias between the three species (Figure 3), and the percentages of genes

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

Table 1. Number of *T. aestivum* transcripts from the NCBI UniGene set (build #60) and full-length cDNA (FL cDNA) sequences retrieved from the TriFLDB database, annotated with a putative function (coverage >50%, identity >45%) in at least one annotation system

Functional annotation systems	Number of annotated transcripts					
	O. sativa		*A. thaliana*		Total[a]	
	NCBI UniGene	FL cDNA	NCBI UniGene	FL cDNA	NCBI UniGene	FL cDNA
MIPS functional classification			12 943	10 864	12 943	10 864
PlantCyc pathway reactions	2193	2106	2093	2208	3067	2911
GOs	13 142	8014	10 444	10 850	16 079	12 279
TAIR *A. thaliana* gene families			4498	3797	4498	3797
MapMan bins	19 248	14 032	13 202	10 897	20 033	14 224
Curated pathways or functions						
Hormone-responsive genes					467	
Ubiquitin-proteasome system					876	
Transcription factors					2891	

[a]Number of transcripts and full-length cDNA sequences annotated with a putative function in at least one model species.

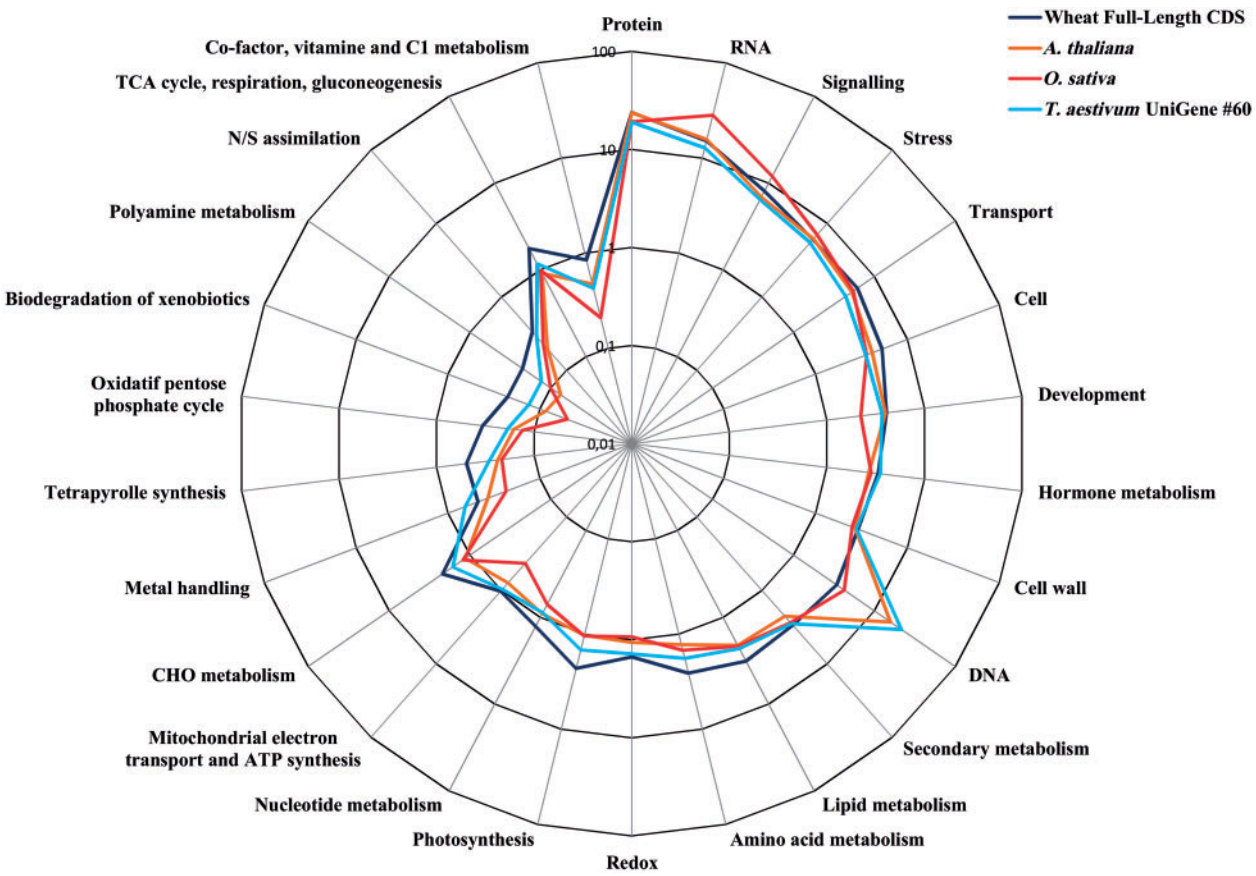

Figure 3. Radar plot (log scale) of the MapMan bin annotations for *A. thaliana*, *O. sativa* and *T. aestivum* UniGene (build #60) and full-length coding sequences. Data are percent of the total number of MapMan bin annotations (Table 1). Similar results were obtained with builds #55, #58 and #59 (data not shown). Some bins have been merged to make the figure clearer.

in the 26 categories for the three species were well correlated (*T. aestivum* versus *A. thaliana*: $r = 0.96$, $P < 0.001$; *T. aestivum* versus *O. sativa*: $r = 0.69$, $P < 0.001$), with no significant bias ($P < 0.001$). The higher correlation found with *A. thaliana* compared with *O. sativa* is mainly because there are fewer annotated transcripts in the DNA bin for *O. sativa* than for *A. thaliana* and *T. aestivum* ($r = 0.90$ for *T. aestivum* versus *O. sativa* when this bin is not considered). For the full-length coding sequences retrieved from TriFLDB and other public databases, the correlations between *T. aestivum* and *A. thaliana* and between *T. aestivum* and *O. sativa* were the same ($r = 0.90$, $P < 0.001$). The pairwise correlations between the four MapMan bin annotations presented were remarkably high, all >0.9 when the DNA bin was omitted. Similar results were obtained for the PlantCyc pathway reactions and GO (data not shown).

Unlike many annotation tools, dbWFA makes it possible to query multiple annotation systems simultaneously. To demonstrate various features of the dbWFA database, some query examples are presented in Box 1, using either the website or the database installed on a local machine.

Identification and Annotation of UniGenes Specifically Expressed During Either the Early or Late Stage of Grain Development

A total of 39 029 transcripts from the UniGene set (build #55) and 1613 transcription factors from the wDBTF database not present in the UniGene set are spotted on the custom *T. aestivum* NimbleGen 40 k microarray (36). Previous studies have shown that 18 140 (44.6%) of these transcripts are expressed during *T. aestivum* grain development (47). In dbWFA, 34–40% (depending on the build) of these transcripts have a putative functional annotation.

Box 1. Query Examples

To demonstrate the usefulness of dbWFA, several biologically relevant queries that can be performed using the current system are presented. In these examples, the UniGene build #55 was used, with coverage and identity thresholds of 50 and 45%, respectively, as recommended by the IWGSC to assign a putative function to a transcript.

Query 1. *Find all* T. aestivum *transcripts likely to have a phytoene synthase activity*

UniGene			Matching sequences		Alignment parameters	
Id number	Representing sequence	Description	Id number	Description	Coverage (%)	Identity (%)
Ta.41960	Ta_S16057905	*T. aestivum* clone wr1.pk0139.g3:fis, full insert mRNA sequence	LOC_OS06G51290	Phytoene synthase, chloroplast precursor, putative, expressed	59.7	81.4
			AT5G17230	Phytoene synthase	58.0	79.6
Ta.66029	Ta_S26027774	FGAS000498 *T. aestivum* FGAS: Library 2 Gate 3? *T. aestivum* cDNA, mRNA sequence	LOC_OS06G51290	phytoene synthase, chloroplast precursor, putative, expressed	55.3	48.9
			AT5G17230	Phytoene synthase	59.7	47.08

The first committed step in the biosynthesis of carotenoids is the condensation of two geranylgeranyl disphosphate molecules by phytoene synthase to produce phytoene, which catalyses a rate-controlling step in the plastid-localized carotenoid pathway (43). We could query the database for the PlantCyc pathway reaction 2.5.1.32 using its web interface. The result of this query is shown in the above table. Two *T. aestivum* transcripts were annotated with a putative phytoene synthase activity. In good agreement with this result, previous studies showed that Poaceae species possess a duplicated phytoene synthase gene (44). A thorough analysis of the two annotated UniGene sequences confirmed that they correspond to the duplicated phytoene synthase gene found in Poaceae. A third phytoene synthase has been isolated in *Z. mays* and *T. aestivum* (45, 46). Although the three *O. sativa* phytoene synthase genes are present in the database, the *T. aestivum* UniGene of this phytoene synthase gene was not found in dbWFA.

The phytoene synthase activity also corresponds to GO:0016767 MapMan bin 16.1.4.1. Searching dbWFA for this GO or MapMan bin yields the same results as above. It is possible to combine several overlying systems (e.g. PlantCyc pathway reaction and GO) in a single MySQL query when the database is installed on a local machine. It is also possible to compare and make the union or intersection of queries using MySQL, depending on the intended outcome.

(continued)

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

Box 1: Continued

Query 2. *Find as much information as possible about a list of transcripts*

UniGene		GO	TAIR	MIPS	PlantCyc	MapMan
Id number	Match					
Ta.41960	**AT5G17230** Phytoene synthase	GO:0009507 GO:0016117 GO:0016767 GO:0046905		01.06.06.13 70.26.03	2.5.1.32 2.5.1.32	16.1.4.1

The efficiency of the database stems from its multiple systems of annotation. The cross-system annotation feature of dbWFA is integrated in the web interface in the 'Transcript(s) annotation' search method. This type of query could be used to obtain information for a list of UniGenes of interest in the different annotation systems integrated in dbWFA. Querying the UniGene set for the first phytoene synthase transcript retrieved in Query 1 yields the annotation shown in the above table. On the web interface, the user can choose to display only the best hit (as in the above table) or the five best hits with percentages of coverage and identity greater than the thresholds set by the user. The user can also choose the systems of annotation to include in the query and the model species. The results redirect the user to the web pages of the different annotation systems, which allows more detailed information to be obtained on the annotation of the list of transcripts of interest.

Query 3. *Find all the transcripts putatively involved in the glycolytic pathway for a transcriptome analysis in MapMan*

Bin code	Name	Identifier	Description	Type
4.1	Glycolysis.cytosolic branch	Ta_S16058223	Similar to UTP–glucose-1-phosphate uridylyltransferase, putative, expressed Coverage: 99.5745%, identity: 92.75%	T
4.1.10	Glycolysis.cytosolic branch.non-phosphorylating glyceraldehyde 3-phosphate dehydrogenase (NPGAP-DH)	Ta_S13048872	Similar to aldehyde dehydrogenase Coverage: 100%, identity: 87.1%	T
4.1.10	Glycolysis.cytosolic branch.non-phosphorylating glyceraldehyde 3-phosphate dehydrogenase (NPGAP-DH)	Ta_S13048873	Similar to aldehyde dehydrogenase Coverage: 100%, identity: 79.23%	T
4.1.11	Glycolysis.cytosolic branch.aldolase	Ta_S15902802	Similar to aldolase superfamily protein Coverage: 50.1873%, identity: 85.07%	T
4.1.11	Glycolysis.cytosolic branch.aldolase	Ta_S17888674	Similar to aldolase superfamily protein Coverage: 88.5475%, identity: 48.91%	T

In the search method 'MapMan mapping file generator', the user can select a metabolic pathway and automatically create a mapping file to visualise the results of transcriptomic experiences performed with the *T. aestivum* custom NimbleGen 40 k microarray using the -omic data viewing and analysing tool MapMan. The glycolytic pathway corresponds to the bin code 4. The first five lines of the table generated by dbWFA for this query are shown above. When the database is installed on a local machine, several pathways could be queried simultaneously to create a custom *T. aestivum* mapping file for MapMan.

T. aestivum grain development comprises several distinct phases, starting with a syncytial then a cellularization phase (ca. 0–100°Cdays after anthesis), followed by a first differentiation phase of active endosperm cell division (ECD), expansion and differentiation (ca. 100–250°Cdays after anthesis), a second differentiation phase when storage polymers rapidly accumulate (ca. 250–750°Cdays after anthesis) and a maturation phase when grain rapidly desiccates (ca. 750–900°Cdays after anthesis) (48, 49). The transitions between these phases are associated with major changes in the grain transcriptome (5, 36, 50, 51) and proteome (52, 53).

To validate the principle underpinning the database and provide another example of the usefulness of dbWFA, we analysed the functional annotation of the transcripts specifically expressed during either the ECD or SPA phase of grain development. We used transcriptome data obtained with the custom *T. aestivum* NimbleGen 40 k microarray for

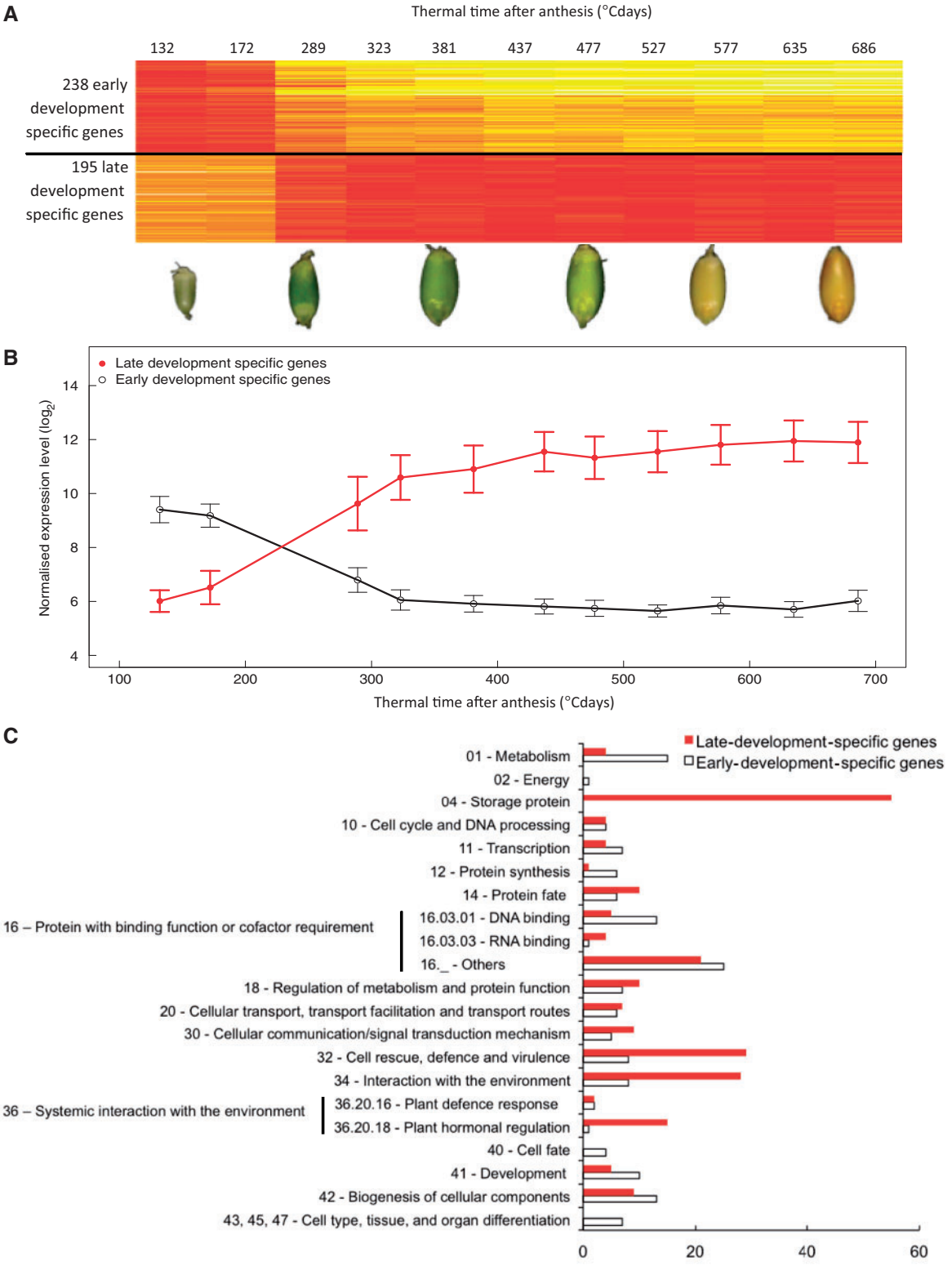

Figure 4. Functional annotation of genes specifically expressed during either the early cell division or late SPA phases of *T. aestivum* grain development. **(A)** Heat map of expression for early- and late-development-specific genes. **(B)** Normalized expression of the early and late development specific gene clusters. Transcripts with normalized expression <7 were not considered to be expressed (i.e. not different from the background noise). Data are medians $\pm$ 1 SD. **(C)** MIPS Functional Categories of genes from both UniGene clusters.

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

the *T. aestivum* cultivar Recital grown under standard conditions in a greenhouse and sampled every 34–117°Cdays between 132 and 686°Cdays after anthesis (35). Transcripts with different patterns of expression were classified with J-Express 2012 software package (54) using Euclidean distance–based k-means clustering. For this analysis, the number of clusters was empirically set at 25 because it allowed us to clearly discriminate between gene expression clusters specific to the ECD and SPA phases of grain development. One cluster of 238 genes contained genes expressed exclusively during ECD stages (Figure 4A and B). Two other clusters contained genes expressed exclusively during SPA stages. The latter two clusters were merged to form a single SPA cluster of 195 genes. dbWFA was then used to retrieve the functional classification of the transcripts from both clusters. The MIPS Functional Classification was used, as it was the most informative and straightforward for comparisons with previous studies.

Using IWGSC-recommended coverage (50%) and identity (45%) percentages, 68 (29%) ECD-specific transcripts and 129 (66%) SPA-specific transcripts were assigned to an MIPS functional category, respectively (Figure 4C). The annotation results were consistent with previous transcriptome (5, 51) and proteome (52, 53) studies of developing *T. aestivum* and *H. vulgare* (55) grain. The functional classifications of the gene clusters were different. Not surprisingly, 12 transcripts involved in cell fate and cell type, tissue differentiation and organ differentiation were specifically expressed during the ECD phase, whereas no SPA phase–specific transcripts were annotated as being in these MIPS functional categories. Also in good agreement with our knowledge of grain development, 55 seed storage protein transcripts were specifically expressed during the SPA phase, while none was found among the annotated ECD-specific genes.

Quantitative differences in the annotation of these two clusters of transcripts were also observed. Several transcripts in the SPA-specific cluster were involved in cell rescue, defence and virulence and in the interaction with the environment. In particular, transcripts involved in plant hormonal regulation were overrepresented in the SPA-specific gene cluster. Transcripts coding for proteins involved in protein synthesis and proteins with metabolic functions were overrepresented in the ECD cluster. These results coincide with previous transcriptome (5, 36, 56) and proteome analyses (53). Finally, we note a substantial difference in the MIPS functional category 'protein with binding function or co-factor requirement', with more transcripts involved in DNA binding in the ECD cluster and more transcripts involved in RNA binding in the SPA cluster.

All these data show great similarity to published results for *T. aestivum* and *H. vulgare*, reflecting the accuracy of the automatic annotation provided by dbWFA. Unlike several other *T. aestivum* transcriptome analyses where a complex process had to be carried out to assign a functional annotation to selected transcripts and/or proteins (5, 57), here only a single request was made to the dbWFA database to retrieve the functional classification of 45% of the transcripts of interest, and 40% of the 40 642 transcripts of the *T. aestivum* NimbleGen 40 k microarray. This percentage of annotated transcripts is similar to that previously reported (38%) for the *T. aestivum* Affymetrix GeneChip® microarray (4).

Outlook

The dbWFA database was created by integrating numerous data sources. As a result, it is a practical source of heterogeneous data for functional annotation of *T. aestivum* transcripts. The website grants access to the most common queries that can be applied to the database, and the freely available MySQL database is a powerful tool for more specific requests. Although further analyses are required to confirm the dbWFA annotation results, the database provides an efficient and fast solution for acquiring a wide range of functional information. cDNA resources are useful to predict exonic regions from genomic sequences; thus, efforts to annotate the UniGene resources will significantly contribute to the analysis of sequence data produced by ongoing *T. aestivum* initiatives and other genome sequencing projects.

The version of dbWFA presented here is operational, but the aim is not to restrict the database to storing *O. sativa* and *A. thaliana* annotations but to expand it to include data from other plant species genomes as their functional annotation becomes more consistent. Integration of InterProScan (58) in the workflow could be a valuable way of augmenting the process. Also, the integration of AFAWE (59) would provide an annotation workflow with different function prediction tools. However, the current version of AFAWE cannot be used independently from its web interface and would thus have to be implemented using the tools called in its workflow, which are available as web services. Finally, the upcoming integration of a BLAST program in the workflow will allow users to annotate their own sequences and will also make dbWFA applicable to other species.

Acknowledgements

The authors thank Dr Etienne Paux and Dr Catherine Feuillet (INRA, UMR1095 GDEC, Clermont-Ferrand, France) for useful discussions and advice, and Mr Sébastien Reboux, Ms Claire Viseux and Mr Michael Alaux (INRA, URGI, Versailles, France) for installing and maintaining the database on the URGI server.

Funding

This work was supported by a PhD grant from the French Ministry for Higher Education and Research to J.V.

Conflict of interest. None declared.

References

1. Feuillet,C. and Eversole,K. (2007) Physical mapping of the wheat genome: a coordinated effort to lay the foundation for genome sequencing and develop tools for breeders. *Isr. J. Plant Sci.*, **55**, 307–313.
2. Feuillet,C., Leach,J.E., Rogers,J. *et al*. (2011) Crop genome sequencing: lessons and rationales. *Trends Plant Sci.*, **16**, 77–88.
3. Leroy,P., Guilhot,N., Sakai,H. *et al*. (2012) TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. *Front. Plant Sci.*, **3**, 1–14.
4. Brenchley,R., Spannagl,M., Pfeifer,M. *et al*. (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**, 705–710.
5. Wan,Y., Poole,R.L., Huttly,A.K. *et al*. (2008) Transcriptome analysis of grain development in hexaploid wheat. *BMC Genomics*, **9**, 121.
6. Pellny,T.K., Lovegrove,A., Freeman,J. *et al*. (2012) Cell walls of developing wheat starchy endosperm: comparison of composition and RNA-Seq transcriptome. *Plant Physiol.*, **158**, 612–627.
7. Köhler,J., Baumbach,J., Taubert,J. *et al*. (2006) Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, **22**, 1383–1390.
8. Lysenko,A., Hindle,M.M., Taubert,J. *et al*. (2009) Data integration for plant genomics–exemplars from the integration of *Arabidopsis thaliana* databases. *Briefings Bioinformatics*, **10**, 676–693.
9. Lazo,G.R., Chao,S., Hummel,D.D. *et al*. (2004) Development of an expressed sequence tag (EST) resource for wheat (*Triticum aestivum* L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16,000-locus bin-delineated map. *Genetics*, **168**, 585–593.
10. Zhang,H., Sreenivasulu,N., Weschke,W. *et al*. (2004) Large-scale analysis of the barley transcriptome based on expressed sequence tags. *Plant J.*, **40**, 276–290.
11. Allouis.,S., Moore.,G., Bellec.,A. *et al*. (2003) Construction and characterisation of a hexaploid wheat (*Triticum aestivum* L.) BAC library from the reference germplasm 'Chinese Spring'. *Cereal Res. Commun.*, **31**, 331–338.
12. Safár,J., Bartos,J., Janda,J. *et al*. (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J.*, **39**, 960–968.
13. Wilkinson,P.A., Winfield,M.O., Barker,G.L.A. *et al*. (2012) CerealsDB 2.0: an integrated resource for plant breeders and scientists. *BMC Bioinformatics*, **13**, 219.
14. Lai,K., Berkman,P.J., Lorenc,M. *et al*. (2012) WheatGenome.info: an integrated database and portal for wheat genome information. *Plant Cell Physiol.*, **53**, e2.
15. Dong,Q., Schlueter,S.D. and Brendel,V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.*, **32**(Database issue), D354–D359.
16. Feuillet,C., Stein,N., Rossini,L. *et al*. (2012) Integrating cereal genomics to support innovation in the Triticeae. *Funct. Integr. Genomics*, **12**, 573–583.
17. Close,T.J., Wanamaker,S., Roose,M.L. *et al*. (2007) HarvEST. *Methods Mol Biol.*, **406**, 161–177.
18. Estill,J.C. and Bennetzen,J.L. (2009) The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. *Plant Methods*, **5**, 8.
19. Mochida,K., Yoshida,T., Sakurai,T. *et al*. (2009) TriFLDB: a database of clustered full-length coding sequences from triticeae with applications to comparative grass genomics. *Plant Physiol.*, **150**, 1135–1146.
20. Manickavelu,A., Kawaura,K., Oishi,K. *et al*. (2012) Comprehensive functional analyses of expressed sequence tags in common wheat (*Triticum aestivum*). *DNA Res.*, **19**, 165–177.
21. The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
22. Yu,J., Hu,S., Wang,J. *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.
23. The International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
24. Schnable,P.S., Ware,D., Fulton,R.S. *et al*. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
25. Schmutz,J., Cannon,S.B., Schlueter,J. *et al*. (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
26. Paterson,A.H., Bowers,J.E., Bruggmann,R. *et al*. (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.
27. The International Brachypodium Initiative. (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, **463**, 763–768.
28. Mayer,K.F.X., Waugh,R., Langridge,P. *et al*. (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.
29. Van Bel,M., Proost,S., Wischnitzki,E. *et al*. (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.*, **158**, 590–600.
30. Dassanayake,M., Oh,D.H., Haas,J.S. *et al*. (2011) The genome of the extremophile crucifer Thellungiella parvula. *Nat. Genet.*, **43**, 913–918.
31. Conesa,A., Götz,S., Garcia-Gomez,J.M. *et al*. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.
32. Lamesch,P., Berardini,T.Z., Li,D. *et al*. (2011) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
33. Ouyang,S., Zhu,W., Hamilton,J. *et al*. (2007) The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.
34. Altschul,S.F., Madden,T.L., Schäffer,A.A. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
35. Romeuf,I., Tessier,D., Dardevet,M. *et al*. (2010) wDBTF: an integrated database resource for studying wheat transcription factor families. *BMC Genomics*, **11**, 185.
36. Capron,D., Mouzeyar,S., Boulaflous,A. *et al*. (2012) Transcriptional profile analysis of E3 ligase and hormone-related genes expressed during wheat grain development. *BMC Plant Biol.*, **12**, 35.
37. Rustenholz,C., Choulet,F., Laugier,C. *et al*. (2011) A 3,000-loci transcription map of chromosome 3B unravels the structural and

Database, Vol. 2013, Article ID bat014, doi:10.1093/database/bat014

functional features of gene islands in hexaploid wheat. *Plant Physiol.*, **157**, 1596–1608.

38. Ashburner,M., Ball,C.A., Blake,J.A. *et al*. (2011) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
39. Zhang,P., Dreher,K., Karthikeyan,A. *et al*. (2010) Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol.*, **153**, 1479–1491.
40. Thimm,O., Bläsing,O., Gibon,Y. *et al*. (2004) Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914–939.
41. Ruepp,A., Zollner,A., Maier,D. *et al*. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.
42. Rhee,S.Y., Beavis,W., Berardini,T.Z. *et al*. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.
43. Cunningham,F.X. and Gantt,E. (1998) Genes and enzymes of carotenoids biosynthesis in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **49**, 557–583.
44. Gallagher,C.E., Matthews,P.D., Li,F. *et al*. (2004) Gene duplication in the carotenoid biosynthetic pathway preceded evolution of the grasses. *Plant Physiol.*, **135**, 1776–1783.
45. Li,F., Vallabhaneni,R. and Wurtzel,E.T. (2008) PSY3, a new member of the phytoene synthase gene family conserved in the poaceae and regulator of abiotic stress-induced root carotenogenesis. *Plant Physiol.*, **146**, 1333–1345.
46. Dibari,B., Murat,F., Chosson,A. *et al*. (2012) Deciphering the genomic structure, function and evolution of carotenogenesis related phytoene synthases in grasses. *BMC Genomics*, **13**, 221.
47. Romeuf,I. (2010) Identification *in silico* des facteurs de transcription du blé tendre (*Triticum aestivum*) et mise en évidence des facteurs de transcription impliqués dans la synthèse des protéines de réserve, *Ph.D. Thesis*. Université Clermont-Ferrand II, Blaise Pascal, Clermont-Ferrand, France, pp. 223.
48. Bennett,M.D., Rao,M.K., Smith,J.B. *et al*. (1975) Cell development in the anther, the ovule, and the young seed of *Triticum aestivum* L. Var. chinese spring. *Philos. T. R. Soc. B*, **266**, 6–81.
49. Evers,T. and Millar,S. (2002) Cereal grain structure and development: some implications for quality. *J. Cereal Sci.*, **36**, 261–284.
50. Drea,S., Leader,D.J., Arnold,B.C. *et al*. (2005) Systematic spatial analysis of gene expression during wheat Caryopsis. *Plant Cell.*, **17**, 2172–2185.
51. Laudencia-Chingcuanco,D.L., Stamova,B.S., You,F.M. *et al*. (2007) Transcriptional profiling of wheat caryopsis development using cDNA microarrays. *Plant Mol. Biol.*, **63**, 651–668.
52. Nadaud,I., Girousse,C., Debiton,C. *et al*. (2010) Proteomic and morphological analysis of early stages of wheat grain development. *Proteomics*, **10**, 2901–2910.
53. Tasleem-Tahir,A., Nadaud,I., Chambon,C. *et al*. (2012) Expression profiling of starchy endosperm metabolic proteins at 21 stages of wheat grain development. *J. Proteome Res.*, **11**, 2754–2773.
54. Dysvik,B. and Jonassen,I. (2001) J-Express: exploring gene expression data using Java. *Bioinformatics*, **17**, 369–370.
55. Sreenivasulu,N., Radchuk,V., Strickert,M. *et al*. (2006) Gene expression patterns reveal tissue-specific signaling networks controlling programmed cell death and ABA- regulated maturation in developing barley seeds. *Plant J.*, **47**, 310–327.
56. Clarke,B.C., Hobbs,M., Skylas,D. *et al*. (2000) Genes active in developing wheat endosperm. *Funct. Integr. Genomics*, **1**, 44–55.
57. Szucs,A., Jäger,K., Jurca,M.E. *et al*. (2010) Histological and microarray analysis of the direct effect of water shortage alone or combined with heat on early grain development in wheat (*Triticum aestivum*). *Physiol Plant.*, **140**, 174–188.
58. Goujon,M., McWilliam,H., Li,W. *et al*. (2010) A new bioinformatics analysis tools framework at EMBL–EBI. *Nucleic Acids Res.*, **38**, W695–W699.
59. Jöcker,A., Hoffmann,F., Groscurth,A. *et al*. (2008) Protein function prediction and annotation in an integrated environment powered by web services (AFAWE). *Bioinformatics*, **24**, 2393–2394.