W148–W151 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm220

Wheat Estimated Transcript Server (WhETS): a tool to provide best estimate of hexaploid wheat transcript sequence

Rowan A. C. Mitchell[1,*], Nathalie Castells-Brooke[1], Jan Taubert[1], Paul J. Verrier[1], David J. Leader[2] and Christopher J. Rawlings[1]

[1]Biomathematics and Bioinformatics Division and [2]Crop Performance and Improvement Division, Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK

Received January 29, 2007; Revised March 23, 2007; Accepted March 28, 2007

ABSTRACT

Wheat biologists face particular problems because of the lack of genomic sequence and the three homoeologous genomes which give rise to three very similar forms for many transcripts. However, over 1.3 million available public-domain Triticeae ESTs (of which ~850 000 are wheat) and the full rice genomic sequence can be used to estimate likely transcript sequences present in any wheat cDNA sample to which PCR primers may then be designed. Wheat Estimated Transcript Server (WhETS) is designed to do this in a convenient form, and to provide information on the number of matching EST and high quality cDNA (hq-cDNA) sequences, tissue distribution and likely intron position inferred from rice. Triticeae EST and hq-cDNA sequences are mapped onto rice loci and stored in a database. The user selects a rice locus (directly or via Arabidopsis) and the matching Triticeae sequences are assembled according to user-defined filter and stringency settings. Assembly is achieved initially with the CAP3 program and then with a single nucleotide polymorphism (SNP)-analysis algorithm designed to separate homoeologues. Alignment of the resulting contigs and singlets against the rice template sequence is then displayed. Sequences and assembly details are available for download in fasta and ace formats, respectively. WhETS is accessible at http://www4.rothamsted.bbsrc.ac.uk/whets.

INTRODUCTION

Wheat is the most widely grown crop in the world with massive importance for human nutrition. However, genomics and DNA sequence analysis in wheat present particular problems. Cultivated wheat (*Triticum aestivum*) is an allohexaploid species with three homoeologous genomes (A, B, D), each comprising seven pairs of chromosomes. All three genomes are very large, so that together they contain about 30× as much DNA as rice and 6× as much as the human genome. Due to the technical difficulties, the complete genome sequence of wheat will not be available for several years at the earliest. However, there is a rich resource of wheat ESTs of which there are ~850 000 and a further ~500 000 from other Triticeae species in dbEST [(1); January 2007]. These can be mapped to the genes of rice as the most closely related fully sequenced genome (2). In this way, all the ESTs derived from the same transcript can be grouped and linked to information on the orthologues in rice and Arabidopsis. This procedure thus facilitates the application of knowledge gained in model species, particularly Arabidopsis, to wheat crops. The aim of Wheat Estimated Transcript Server (WhETS) is to flexibly allow the user to assemble Triticeae ESTs mapped to rice genes in this way and provide access to the results in a convenient form.

A common way to exploit ESTs is to use the pre-existing assemblies such as Unigene (3). However, because WhETS assembles related sequences in real time, the user can adjust the set of ESTs to be used, alter the stringency setting and view the affect of the changes on the assembly. Also, by anchoring the ESTs to rice loci, non-contiguous ESTs representing the same genes are

*To whom correspondence should be addressed. Tel: +44 (0)1582 763133; Fax: +44 (0)1582 763010; Email: rowan.mitchell@bbsrc.ac.uk
Present address:
David J. Leader, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

automatically treated a part of the same set. The three very similar homoeologues of wheat genes are frequently all expressed (4), but assembly programs do not normally separate these so they are grouped together in contigs. These homoeologous sequences are best identified by analysis of shared SNPs such as can be achieved with SNPserver (5) which uses autoSNP (6) algorithms to separate alleles or homoeologues. A similar approach is included in WhETS to provide a best estimate of homoeologue-specific sequence. Aligning the Triticeae ESTs to rice has the additional advantage of being able to infer likely intron position which can be used to derive allelic markers, an approach taken in the USDA wheat SNP database (http://wheat.pw.usda.gov/SNP/new/index.shtml). Part of the aim of WhETS is therefore to bring together the useful features of Unigene, SNPserver and USDA SNP database into one site tailored specifically for wheat transcripts. However, it also has features unavailable elsewhere, such as the ability to display Triticeae EST distribution corresponding to a set of rice loci and the option to filter sequences according to library source tissue.

DESCRIPTION

Database

WhETS has a relational database containing sequences and annotation for all Triticeae EST and high quality cDNA (hq-cDNA) sequences from dbEST, coding sequences, annotation and intron positions for rice from The Institute for Genome Research (TIGR) rice pseudo molecule release 4 (7), and annotation for each locus from The Arabidopsis Information Resource (TAIR) version 6 (8) (Figure 1). EST sequences are first masked for vector contamination using the cross_match program (9). The WhETS database also contains the results of a blast (10) similarity searches: blastp of all the TIGR rice protein sequences against all the TAIR Arabidopsis proteins, and blastn of the Triticeae ESTs against the TIGR rice CDS. These tables contain the top scoring hits and any lower scoring hits with longer aligned regions, thus defining many-to-many relationships between Arabidopsis and rice genes and between rice genes and Triticeae sequences. The database is updated weekly by automated scripts which compare the contents with Triticeae

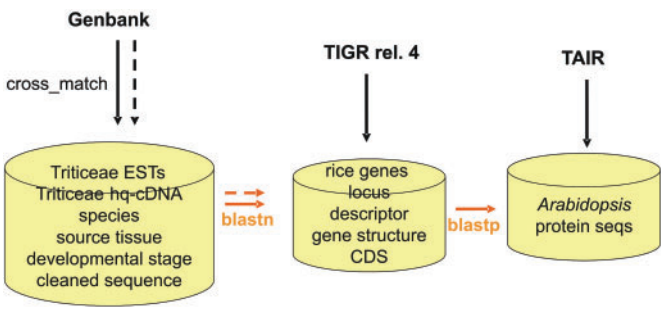

Figure 1. WhETS database preparation steps. Dashed arrows indicate steps which are repeated in automatic weekly updates.

entries in GenBank using Entrez utilities (11). Any missing entries are downloaded and any extra ones deleted. New sequences are subjected to a blastn search against the rice sequences and the sequences, and blast results are added to the WhETS database (Figure 1).

Real-time operation

The main part of WhETS requires TIGR rice loci identifiers. Users can start directly by supplying these as input, or they can start with a set of Arabidopsis AGI numbers or Triticeae accession numbers. WhETS will then retrieve all the matching rice loci for these. The user can then select filter settings for species, tissue and sequence type (EST or hq-cDNA). WhETS will then display the number and accessions of all the matching Triticeae sequences for each locus. The user can then select the locus for which they wish to obtain sequences for in the main part of WhETS.

When a single rice locus is selected, the user can again filter for species, tissue and sequence type. The sequences which pass this filter are assembled using the CAP3 program (12), and the resulting contigs are passed to an algorithm which analyses shared SNPs. If the contigs are found to contain groups which share more SNPs (i.e. base differences from the consensus) than a user-defined cut-off (default five SNPs per kb), these are split off into separate contigs. The CAP3 step tends to assemble paralogues which match the same rice locus into separate contigs, whereas the SNP analysis step is designed to separate homoeologues. However, by selecting higher stringency the user may also separate allelic forms. Conversely, in situations where there are relatively few ESTs it can be useful to assemble the sequences from wheat and related species with low stringency. WhETS also assembles the hq-cDNA sequences where present using a much higher base quality setting for the CAP3 program than used for ESTs. This has the effect that the consensus sequence of any contig will normally be the same as any hq-cDNA within it.

After assembly, the rice CDS is aligned to the contigs' consensus and singlet sequences with blast and the results displayed using a modified version of a Perl script from the Korf *et al.* study (13). For contigs, links are supplied which open windows detailing all species, tissue, sequence type and cultivar of the constituent sequences. Singlets link out to the original NCBI entry. The main output for user downloading is a fasta file containing the rice template CDS, contigs' consensus and singlet sequences. Additional details, such as intron positions are supplied in the descriptor fields of this file. Also available are other files, such as ace format files for each of the contigs containing all the information on the constituent sequences and their alignment, and a spreadsheet-compatible file containing details of all SNPs used to split contigs.

WhETS is implemented in MySQL (http://www.mysql.com/) and Perl using some Bioperl (14) modules. More details on allocation of blast hits within the WhETS database, strand of EST

used and the algorithm for separating contigs into putative homoeologues are available in the Supplementary Data.

EXAMPLE OUTPUT

To test that WhETS correctly separates homoeologues, we examined the well-characterized WAXY locus, which encodes granule-bound starch synthase I. The three homoeologous forms are all sequenced, as are several allelic variants of these. As there are only a total of 2 715 wheat hq-cDNA sequences available, the normal use of WhETS is only with ESTs. We, therefore, ran WhETS with the orthologous rice locus Os06g04200 setting the filter to use ESTs and wheat sequences only. Figure 2 shows the output and how the resulting contigs match with the known homoeologues. From ESTs alone, WhETS correctly identifies the homoeologues and indicates the existence of a splice variant of the B homoeologue with a deletion in its 5′ UTR. Also shown (Figure 3) is the additional window detailing constituent sequences of one of the contigs.

CONCLUSION

WhETS is designed to be a practical tool for wheat biologists to rapidly get the best estimate of transcript sequence for a target gene, supplemented with information on tissue distribution and likely gene structure. It is particularly aimed at producing wheat sequences from which to design PCR primers for cDNA templates.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

Rothamsted Research receives grant-aided support from the Biotechnology and Biological Sciences Research

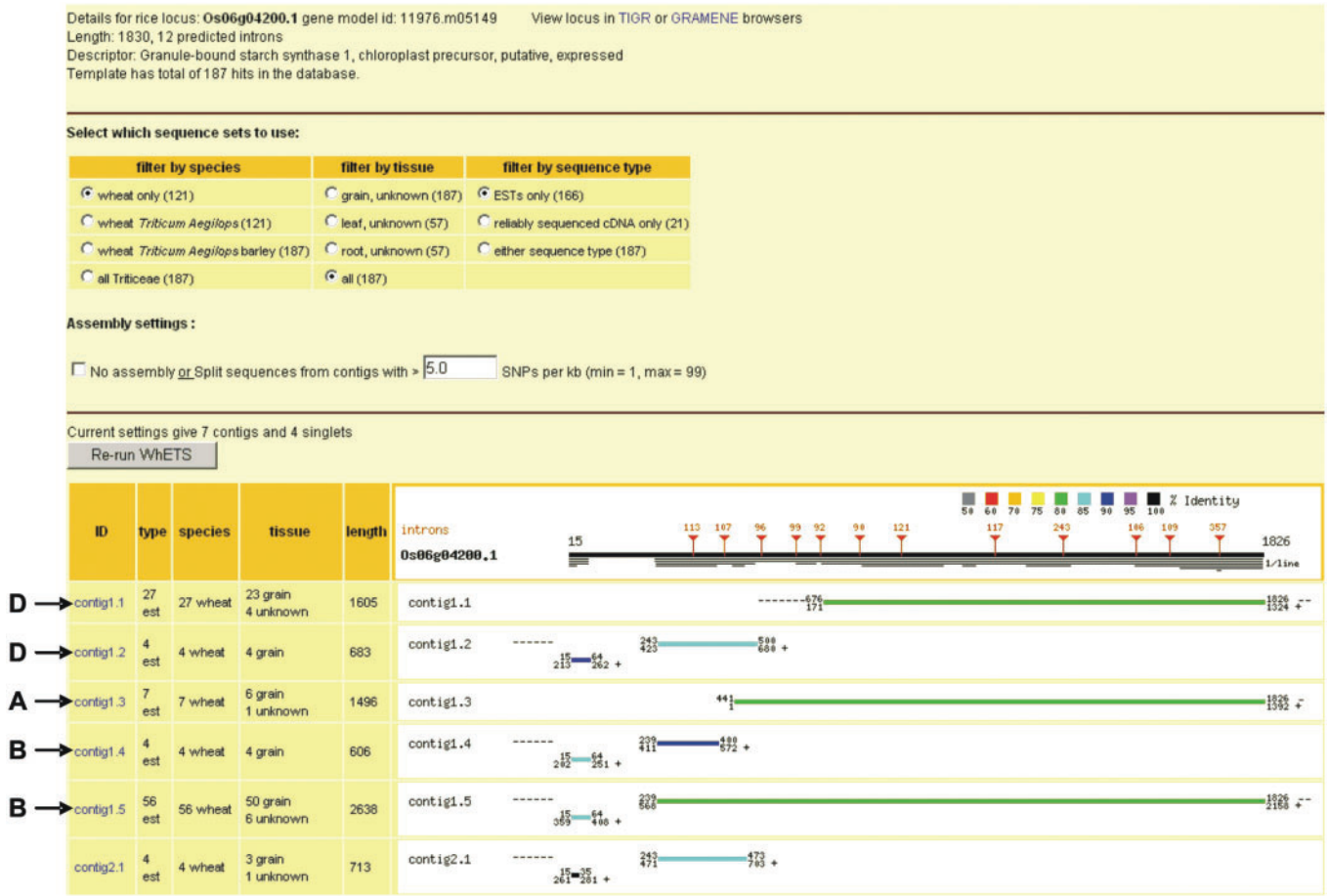

Figure 2. Output from WhETS for Os06g04200.1. The black line at the top corresponds to the rice gene CDS with intron position and size indicated as red vertical lines with triangles. Thin horizontal lines below this indicate the coverage of hits from the Triticeae sequences. The rows below show these hits, with blast HSPs for contigs and singlets shown as lines coloured according to percentage identity, and coordinates aligned to the rice template. The CAP3 step gives three contigs; one of these (contig 1) is then divided into five new contigs by the SNP analysis step (contigs 1.1, 1.2, etc.) The genome of origin (A, B, D) has been added to the screenshot according to 100% identity matches of the contig consensus to the known-homoeologue transcript sequences (exons of accessions AB019622, AB019623 and AB019624). Contigs 1.1 and 1.2 are not combined because of a lack of substantial overlap. Contigs 1.4 and 1.5 appear to be splice variants with an indel in the 5′UTR. Contigs 2 and 3 appear quite different and may be paralogues.

Os06g04200.1/wheat_all_est/lim50 contig1.1

ID	type	species	tissue	length	cultivar
AL810563	ests	Triticum aestivum	endosperm	329	Mercia
BE414707	ests	Triticum aestivum	endosperm	615	Wyuna
BQ607136	ests	Triticum aestivum	endosperm	479	Wyuna
BQ244862	ests	Triticum aestivum	developing seeds	325	Glenlea
BE402015	ests	Triticum aestivum	endosperm	580	Wyuna
BQ607427	ests	Triticum aestivum	endosperm	580	Wyuna
BQ238877	ests	Triticum aestivum	developing seeds	480	Glenlea
AW448881	ests	Triticum aestivum		634	Wyuna
AL814667	ests	Triticum aestivum	endosperm	523	Mercia
CA715510	(rc) ests	Triticum aestivum	kernel	442	
CA741778	ests	Triticum aestivum	anthers	615	
BE423625	ests	Triticum aestivum	Endosperm	522	Cheyenne
BQ246353	ests	Triticum aestivum	developing seeds	702	Glenlea
BE606874	ests	Triticum aestivum	Spike	383	Chinese Spring
CD939860	ests	Triticum aestivum	ovary	650	recital
BQ245812	ests	Triticum aestivum	developing seeds	572	Glenlea
CD905280	ests	Triticum aestivum	grain (468 degrees per day after pollination)	497	recital
AL814554	ests	Triticum aestivum	endosperm	595	Mercia
BJ235041	ests	Triticum aestivum	seed DPA10	601	Chinese Spring
CD453892	(rc) ests	Triticum aestivum	Spike	590	Chinese Spring
CD934500	(rc) ests	Triticum aestivum	ovary	603	recital
BQ237731	(rc) ests	Triticum aestivum	developing seeds	599	Glenlea
CA707163	ests	Triticum aestivum	kernel	438	
BQ242088	(rc) ests	Triticum aestivum	developing seeds	580	Glenlea
BQ244546	(rc) ests	Triticum aestivum	developing seeds	557	Glenlea
BQ243594	(rc) ests	Triticum aestivum	developing seeds	553	Glenlea
BQ235970	(rc) ests	Triticum aestivum	developing seeds	474	Glenlea

Figure 3. Display window showing details for a contig which is opened by clicking on contig 1.1 link in Figure 2.

Council of the United Kingdom. Funding to pay the Open Access publication charges for this article was provided by Rothamsted Research.

Conflict of interest statement. None declared.

REFERENCES

1. Boguski,M.S., Lowe,T.M.J. and Tolstoshev,C.M. (1993) Dbest – Database for expressed sequence tags. *Nature Genet*., **4**, 332–333.
2. Matsumoto,T., Wu,J.Z., Kanamori,H., Katayose,Y., Fujisawa,M., Namiki,N., Mizuno,H., Yamamoto,K., Antonio,B.A. *et al*. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
3. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. *et al*. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res*., **31**, 28–33.
4. Mochida,K., Yamazaki,Y. and Ogihara,Y. (2004) Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. *Mol. Genet. Genomics*, **270**, 371–377.
5. Savage,D., Batley,J., Erwin,T., Logan,E., Love,C.G., Lim,G.A.C., Mongin,E., Barker,G., Spangenberg,G.C. *et al*. (2005) SNPServer: a real-time SNP discovery tool. *Nucleic Acids Res*., **33**, W493–W495.
6. Barker,G., Batley,J., O'Sullivan,H., Edwards,K.J. and Edwards,D. (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics*, **19**, 421–422.
7. Ouyang,S., Zhu,W., Hamilton,J., Lin,H., Campbell,M., Childs,K., Thibaud-Nissen,F., Malek,R.L., Lee,Y. *et al*. (2007) The TIGR rice genome annotation resource: improvements and new features. *Nucleic Acids Res*., **35**, D883–D887.
8. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G.H., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G. *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.
9. Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*., **8**, 175–185.
10. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
11. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) GenBank. *Nucleic Acids Res*., **34**, D16–D20.
12. Huang,X.Q. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res*., **9**, 868–877.
13. Korf,I., Yandell,M. and Bedell,J.A. (2003) *BLAST* O'Reilly, Sebastopol, USA.
14. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G.R., Korf,I. *et al*. (2002) The bioperl toolkit: perl modules for the life sciences. *Genome Res*., **12**, 1611–1618.