ISSN 1479-2621

Plant Genetic Resources: Characterization and Utilization (2011) **9**(2); 181–184
doi:10.1017/S1479262111000281

# Next-Gen sequencing of the transcriptome of triticale

Y. Xu[1], C. Badea[1], F. Tran[2], M. Frick[1], D. Schneiderman[2], L. Robert[2], L. Harris[2], D. Thomas[1], N. Tinker[2], D. Gaudet[1] and A. Laroche[1]*

[1]*Agriculture and Agri-Food Canada, Research Centre, 5403 1st Avenue S., Lethbridge, AB, Canada T1J 4B1 and* [2]*Agriculture and Agri-Food Canada, ECORC, 960 Carling Avenue, Ottawa, ON, Canada K1A 0C6*

## Abstract

Triticale possesses favourable agronomic attributes originating from both its wheat and rye progenitors, including high grain and biomass yields. Triticale, primarily used as animal feed in North America, is an excellent candidate for production of industrial bio-products. Little is known about the coordination of gene expression of rye and wheat genomes in this intergeneric hybrid, but significant DNA losses from the parental genomes have been reported. To clarify the regulation of gene expression in triticale, we carried out 454 sequencing of cDNAs obtained from root, leaf, stem and floral tissues in different lines of triticale and rye exhibiting different phenotypes and assembled reads into contigs. Related to the data assembly were the absence of reference genomes and the paucity of rye sequences in GenBank or other public databases. Consequently, we have sequenced cDNA libraries from roots, seedlings, leaves, floral tissues and immature seeds to facilitate the identification of triticale sequences originating from rye. To further characterize the wheat-derived cDNAs, we also developed a database close to 25,000 non-redundant full-length wheat coding sequence genes, based on existing databases and contigs that were verified against protein sequences from the grass genomes of *Brachypodium distachyon*, rice, sorghum and maize.

**Keywords:** next-generation sequencing; transcriptome; Triticale

## Introduction

Triticale (×*Triticosecale* Wittm.) is an intergeneric hybrid between wheat species (*Triticum* ssp. AA, AABB and AABBDD) and rye (*Secale cereale*, RR). Triticale possesses favourable agronomic attributes originating from both its wheat and rye progenitors, including high grain and biomass yields. Although significant DNA losses from the parental genomes have been reported (Ma and Gustafson, 2008), little is known about the coordination of gene expression of rye and wheat genomes in this intergeneric hybrid. So far, public databases (e.g. GenBank) are reporting about 14,000 Expressed Sequence Tag (EST) or DNA sequences of rye and a few dozen sequences for triticale.

Next-generation sequencing provides an efficient tool to address biological questions at the transcriptomic and genomic levels. RNA-Seq for transcriptomics (Wang *et al.*, 2009) has been successfully employed in crop plants (Zenoni *et al.*, 2010; Zhang *et al.*, 2010).

We are reporting on the utilisation of the Roche 454 sequencing technology to investigate the transcriptome of triticale and rye in different tissues and at different developmental stages. This preliminary analysis reports on gene expression and identification of triticale- and rye-assembled genes and on the development of an enhanced full-length (FL) cDNA database of wheat to facilitate our analysis.

## Materials and methods

Tissues including leaf, root, stem and different reproductive organs (stigma, anthers, pollen and immature heads)

*Corresponding author. E-mail: andre.laroche@agr.gc.ca

https://doi.org/10.1017/S1479262111000281 Published online by Cambridge University Press

from genotypes of rye (Prima and Vacaria) and triticale (AC Certa, hollow stemmed; Triticale 797 and Triticale 1308, both solid stemmed) sampled at different stages of development were used in this study. AC Certa seedling tissues were also exposed to water stress.

RNA-Seq was carried out. Briefly, total RNA (Trizol (Invitrogen) followed by Qiagen RNeasy midi purification) was extracted from the different tissues. PolyA$^+$ mRNA using Poly (A) Purist™ Kit (Ambion, Inc) was purified from 0.6 mg of total RNA followed by cDNA library synthesis. Five micrograms of double stranded cDNA was utilized for 454 sequencing. Eleven triticale and eight rye libraries were sequenced using the 454 GS FLX Titanium (Roche) technology. Different publically available and proprietary software programs were used in the analysis of 454 sequencing results and included: BLAST (ftp://ftp.ncbi.nih.gov/blast/; Altschul *et al.*, 1997), TGICL (http://compbio.dfci.harvard.edu/tgi/software/; Pertea *et al.*, 2003), CAP3 (Huang and Madan, 1999), SeqClean (http://compbio.dfci.harvard.edu/tgi/software/), OrthoMCL 1.4 (http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi), CD-HIT-EST (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi), BioPerl module Bio::SeqIO and Bio::SearchIO (http://www.bioperl.org/wiki/Main_Page) and DNASTAR SeqMan and NGEN (DNASTAR, Madison, WI, USA) for *de novo* assembly.

## Results and discussion

### *Assembling sequences reflects the genome divergence of triticale and rye*

Sequencing results from eleven triticale libraries generated 3,310,375 reads with an average length 320 bp, while results from eight rye libraries yielded 3,124,641 reads with an average length 345 bp. *De novo* assembly of these two datasets with parameters set at 92% identity and minimum match of 30 nt yielded 162,686 contigs from triticale, which was 35% higher than the number of contigs obtained from rye, 120,416. Interestingly, the number of long contigs above 2 kb from rye, 4657, was almost double than those from triticale, 2774. Given the similar number of input sequences for assembly of rye and triticale, it was not surprising to identify a smaller number of contigs above 2 kb in triticale due to the divergence between the three subgenomes (AABBRR) of triticale (Chalupska *et al.*, 2008) and the stringent parameters used for both the rye and triticale contigs assembly.

To further investigate expressed sequence divergence between triticale and rye, we compared these contigs to the *Brachypodium* proteome using BLASTX. Triticale contigs matched 17,915 (70.3%) *Brachypodium* proteins, whereas 19,142 (75.1%) *Brachypodium* proteins were recognized by rye contigs. The triticale and rye sequences displayed homology to 15,904 *Brachypodium* proteins, while triticale- and rye-assembled sequences identified 2011 and 3777 specific proteins, respectively. These results clearly show that a substantial proportion of rye sequences are not expressed in triticale.

We also combined all the rye and triticale 454 reads together and conducted a second *de novo* assembly utilising the same parameters used for the individual rye and triticale assembly. When we parsed the contig makeup, we found that most contigs were made up of sequences from only one species, triticale or rye (Fig. 1), thus clearly indicating the genome origin of the majority of triticale transcripts.

Sequence variation between rye (RR) and triticale (AABBRR) was observed as exemplified in Fig. 2. As indicated by the arrows, two diagnostic nucleotides between rye and triticale were identified. In triticale, five and possibly six reads out of the 14 sequences were clearly related to the rye genome. This can also provide information on relative expression of homeologous sequences from wheat and rye in triticale when applied to datasets from each individual cDNA library.

## Enhancing the wheat FLcDNA dataset for analysis of triticale and other small grain cereals

In order to compare transcriptional units of rye (RR), triticale (AABBRR) and common wheat (AABBDD), we built an enhanced reference FLcDNA dataset starting

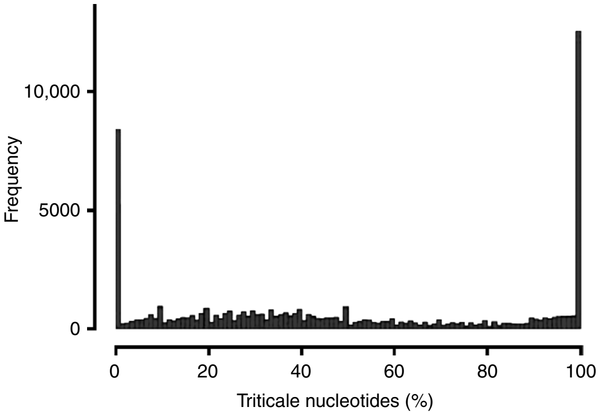

**Fig. 1.** Distribution of sequences originating from triticale and rye libraries in the assembly of contigs. For any given contig on the *x*-axis, we have evaluated the contribution of the number of sequences from rye (Nr) and triticale (Nt) using the following formula Nt/(Nt + Nr) × 100. The value at '0' represents contigs with 100% rye sequences, while the value at '100' represents contigs with 100% triticale sequences. This could be used to estimate the divergence between triticale and rye sequences within each individual contig.

https://doi.org/10.1017/S1479262111000281 Published online by Cambridge University Press

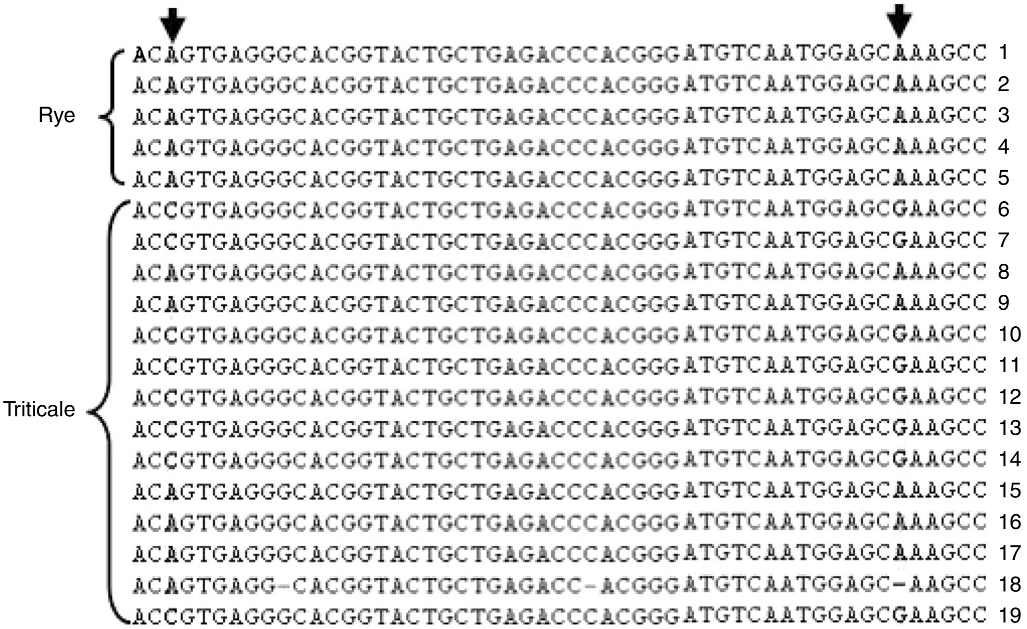
Rye
Triticale
ACAGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCAAAGCC 1
ACAGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCAAAGCC 2
ACAGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCAAAGCC 3
ACAGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCAAAGCC 4
ACAGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCAAAGCC 5
ACCGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCGAAGCC 6
ACCGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCGAAGCC 7
ACAGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCAAAGCC 8
ACAGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCAAAGCC 9
ACCGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCGAAGCC 10
ACCGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCGAAGCC 11
ACCGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCGAAGCC 12
ACCGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCGAAGCC 13
ACCGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCGAAGCC 14
ACAGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCAAAGCC 15
ACAGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCAAAGCC 16
ACAGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCAAAGCC 17
ACAGTGAGG-CACGGTACTGCTGAGACC-ACGGGATGTCAATGGAGC-AAGCC 18
ACCGTGAGGGCACGGTACTGCTGAGACCCACGGGATGTCAATGGAGCGAAGCC 19

**Fig. 2.** Nucleotide variation to distinguish ESTs originating from rye and wheat genomes in triticale. The upper panel represents the nucleotide composition from rye, while the lower panel represents the nucleotide composition detected in triticale. The sequence corresponding to rye can be found in triticale in lines 8, 9, 15, 16, and 17.

with 1,067,304 wheat EST sequences available in GenBank. Firstly, we optimized the assembly pipeline by introducing a 'noise site correction' step for the individual ESTs and an iterative assembly step based on the widely used EST assembly program TGICL (Pertea *et al.*, 2003). More specifically, we used a first round CAP3 assembly with overlap of 40 nt and per cent identity of 80% to remove the 'noise' nt from the aligned sequences. Then, a second round of CAP3 assembly, still with an overlap of 40 nt but with a per cent identity increased to 95%. Under these conditions, the assembly of over 1 million wheat EST dataset yielded a total of 79,034 consensus sequences. The FLcDNA prediction pipeline guided by four known grass genomes (*Brachypodium*, rice, maize and sorghum) identified 21,756 FL wheat cDNAs. The 21,002 publically available FLcDNAs from TriFLDB (11,877; Mochida *et al.*, 2009) and GenBank (9125) were combined and treated with 'Cd-hit-est' to delete the redundancy at 95% identical level to yield 12,715 non-redundant (nr) sequences. This number was almost doubled to 24,789 nr wheat FLcDNAs when our FLcDNA-EST dataset was amalgamated under the same parameters. A comparison to the *Brachypodium* proteome suggests that the expanded nr wheat FLcDNA dataset covers the majority of the proteins of this reference species.

To evaluate the usefulness of the nr wheat FLcDNA dataset, we mapped 384,470 triticale 454 ESTs using MegaBLAST. More than 66% of our 454 reads could be aligned to our enhanced nr wheat FLcDNAs dataset. This dataset will also be very valuable for any wheat RNA-Seq project.

We have assembled *de novo* rye and triticale 454 datasets, distinguished rye from wheat sequences and developed an enhanced nr wheat FLcDNA dataset of almost 24,800 distinct elements.

## Acknowledgements

Funding from the AAFC-ABIP program to The Cellulosic Biofuel Network (159) and the Canadian Triticale Biorefinery Initiative (227) projects and from Alberta Energy–Genome Alberta is greatly appreciated.

## References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25: 3389–3402.

Chalupska D, Lee HY, Faris JD, Evrard A, Chalhoub B, Haselkorn R and Gornicki P (2008) *Acc* homoeoloci and the evolution of wheat genomes. *The Proceeding of the National Academy of Sciences USA* 105: 9691–9696.

Huang X and Madan A (1999) CAP3: a DNA sequence assembly program. *Genome Research* 9: 868–877.

Ma XF and Gustafson JP (2008) Allopolyploidization-accommodated genomic sequence changes in triticale. *Annals of Botany* 101: 825–832.

Mochida K, Yoshida T, Sakurai T, Ogihara Y and Shinozaki K (2009) TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. *Plant Physiology* 150: 1135–1146.

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J and Quackenbush J (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19: 651–652.

Wang Z, Gerstein M and Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* 10: 57–63.

Zenoni S, Ferrarini A, Giacomelli E, Xumerle L, Fasoli M, Malerba G, Bellin D, Pezzotti M and Delledonne M (2010) Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-Seq. *Plant Physiology* 152: 1787–1795.

Zhang G, Guo G, Hu X, Zhang Y, Li Q, Li R, Zhuang R, Lu Z, He Z, Fang X, Chen L, Tian W, Tao Y, Kristiansen K, Zhang X, Li S, Yang H, Wang J and Wang J (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Research* 20: 646–654.

https://doi.org/10.1017/S1479262111000281 Published online by Cambridge University Press