A comparative phylogenetic analysis of hexaploid wheat evolution

ABSTRACT

Bread wheat (Triticum aestivum L.) hexaploid genome is a result of several multiple hybridization events between three diploid progenitor species, comprising three individual subgenomes: A, B, and D. To track the events that have shaped the modern wheat genome, a phylogenetic-based in silico model was developed. The model is based on a comparative analysis between each of the three subgenomes and its progenitor: the subgenomes A and D were compared respectively to Triticum urartu and Aegilops tauschii, while for the still disputable B subgenome predecessor the model plant Brachypodium distachyon was used. The phylogenetic trees for this analysis were obtained from Ensembl Plants. All analyses were done with custom-developed Python scripts. The dataset shows almost equal gene distribution between the three subgenomes: A – 31%, B – 37.3%, and D – 31.7%. The evolutionary relationships were traced using a Species Overlap algorithm, and 54 807 groups of homologs involving wheat genes were detected, including 35 933 orthologous and 18 874 paralogous relationships. The homeologous relations between the three subgenomes were included within the paralogs. 77% of the orthologous groups involved Triticum urartu, 79% - Aegilops tauschii and 64% - Brachypodium distachyon. The paralogs between Triticum urartu and Triticum aestivum were 68%, between Aegilops tauschii and Triticum aestivum, – 71%, and for the Brachypodium distachyon – 42%. Comparative phylogenetic analysis allows identification of the closest homolog and probable predecessor for majority of the available wheat genes. The distribution of genes, orthologs and paralogs also implies an opportunity for subsequent functional prediction and other related analyses.

Key words: hexaploid wheat, evolution, phylogeny, in silico comparative analysis

Introduction

Cereals account for over 50% of the total crop production in the world (http://www.fao.org/), as their seeds are one of crucial importance as renewable resources for food, feed and industrial raw materials (Mochida and Shinozaki, 2013).

Bread wheat (Triticum aestivum) is one of the first cultivated grain crops and ranks first among the harvested crops, having a vital significance for the livelihood of mankind in global sense. The genome of the bread wheat is a hexaploid (6n), with a length of approximately 17 Gbp (almost 6 times as large as the human genome). The genome is composed of three closely related sub-genomes, conditionally named A, B, and C. This complicated structure is a result of two independent evolutionary events of hybridization for which elucidation exist several hypotheses (Bolser et al., 2014; IWGSC 2014; Pfeifer et al, 2014).

It is assumed that the formation of the wheat species Triticum and Aegilops from the other cereals with a common origin in two separate genomes A and B, began about 6.5 million years ago (Figure 1). The first hybridization occurs about 5.5 million years precisely between these two genomes
A and B, and constitutes the formation of a new genome – D, through homoploid hybridization (Marcussen et al., 2014; Kilian et al., 2009).

Second important hybridization event in the evolutionary formation of the contemporary wheat genome has occurred between the close predecessor relatives *Aegilops speltoides* (BB) and *Triticum urartu* (AA), which underwent poliploidization with a wild wheat type *Triticum turgidum* (AABB), which is a lotetraploid (Marcussen et al., 2014).

Finally the modern bread wheat comes from alopoliploidization between these species and wild diploid *Aegilops tauschii*, who has formed DD genome. This last hybridization stage, has held 10 - 8000 years ago, leading to the evolutionary speciation of *Triticum aestivum*, with hexaploid genome structure consisting of of three subgenomes AABBDD (Marcussen et al., 2014; Kilian et al., 2009; Jia et al., 2013; Ling et al., 2013; Murat et al., 2014).

Aside with all evolutionary processes in bread wheat development, between the constituent genomes exist evidences for dynamic loss and gain of genes, as well as duplications, due to different origins (Moore et al., 2005).

Gene duplications in wheat in the context of functionality can be considered as potential sources of adaptation and diversification (Adams et al., 2005). In our study the discovery of gene duplications was done using *in silico* comparative genome analysis, which is an essential approach in clarifying the organization, function and evolution of large polyploid genomes like that of bread wheat, and certainly would help find genes associated with important agronomic traits.

Duplicated genes that remain in the genome may retain its current function, partially or completely may lose its function ("loss" function) or may obtain partially or completely new features ("gain" function). The opposite process of the "loss" of genes results in a decrease of the copies by other mutational events under a so-called "fractionation" such as: deletions, insertions, randomization by substitution of base pairs, overwriting simple repeats and other (Langham et al., 2004, Adams et al., 2005; Moore et al., 2005).

Based on the described evolutionary events different relations between genes depending on their origin have been formed. Genes derived from a common ancestor have similar sequence both with him and with each other. This phenomenon is called homology, and such genes - homologs. Homologous genes derived as a result of speciation are called orthologs. They usually have similar function, but are located in different species. Identification of orthologs is crucial for the reliable prediction of gene functions in new genome sequences.

![Figure 1. Model of the evolutionary history of bread wheat (Triticum aestivum, AABBDD).](image-url)
RESEARCH ARTICLE

Genes derived as a result of duplication event are called paralogs and usually have a different function. Duplication arising before speciation leads to external paralogs. Foreign paralogs, in turn, are divided into interspecific and intraspecific. When speciation precedes duplications internal paralogs have been formed. These internal paralogs belong to only one species (Barton et al., 2007).

There is another type of paralogs called homeologs - that are homologous genes located in chromosome pairs from different sub-genomes resulting overall genomic duplication and polyploidization. Homeology plays a crucial role in the genome of bread wheat and is the main reason for its difficult deciphering. In the genome of hexaploid wheat the closely related (cognate) genes, divided into the three sub-genomes are homeologs between themselves each other (Junhua et al., 2013; Benchley et al., 2012; Krasileva et al., 2013).

Principal method for studying genome evolution is comparative genomic analysis. Its main objective is to discover new structural variants and mutations defining major adaptation properties. This type of analysis also simultaneously compares the studied organism with the genomes of other species – in our work, with plants of the genus Triticeae, also of vital economic importance to agricultural production for food and feed.

An integral part of the comparative genomic analysis using bioinformatics tools is to compare the evolutionary relationships and to identify essential events for prediction of gene functions associated with important agronomic traits, implemented in modern breeding practices intensively using genomic information (Krasileva et al., 2013).

Goal and tasks

Our work is an attempt to trace the evolutionary events leading to the formation of modern genome of bread wheat (Triticum aestivum L.) as a result of hybridization of the genomes of its predecessors. The goal of the study is to track the dynamics of genes during an observed time horizon of the evolution events as well as to consider which of them have been preserved and which have dropped out in different stages. The whole work is based on in silico methods for analysis of phylogenetic trees of genes and comparison of the modern hexaploid wheat genome with the genomes of two already known diploid predecessors - Triticum urartu, Aegilops tauschii, and the model plant Brachypodium distachyon.

To achieve the goal of the study the following tasks were formulated:

1. Choosing of right database with sufficient information about the evolutionary links between wheat and the studied predecessors at the level of genes;
2. Normalization of data – eliminate any potential errors, inaccuracies, extraneous information, etc.;
3. Tracking the evolutionary relationships between genes of the studied species;
4. Mapping of the genes to the diploid progenitors and to the corresponding sub-genomes of the hexaploid wheat.

Materials and Methods

The data used in the study is from the database Ensembl Plants (http://plants.ensembl.org/index.html), which is a subsection of the resource for comparative genomics Ensembl Genomes (http://ensembl.genomes.org/), part of the Ensembl (www.ensembl.org). Ensembl Genomes is a suit with software tools providing also various options for analysis such as different BLAST-like methods (Basic Local Alignment Search Tool - Altschul et al., 1990) for sequence alignment. Ensembl Plants is a rapid developing module from Ensembl which handle genome information for all sequenced plant species (Bolser, 2014). Ensembl Plants has updates 4-5 times per year, with the updating of Ensembl, which uses the same synchronization software. In this work was used release 27 of the Ensembl Plants.

In this line for the purposes of our work originally was extracted general information about the evolutionary links between all species recorded in Ensembl Plants. In studies aiming at revelation of the evolutionary events and their interactions between major type of presentation and events description in classes of organisms is given by phylogenetic tree (Barton et al., 2007).

Information from release 27 including dataset with gene phylogenetic trees is given in Newick format. Newick is a format standard, which is included in most software packages for representing and construction of phylogenetic trees that can be presented in series of nested parentheses enclosing the names of nodes separated by commas.

The strategy of the analysis in our study was structured in several consecutive stages:

a) Extract general information in a datasets comprising number of species, number of trees, number of genes and genes number within the trees.
b) Indexing of the phylogenetic trees in order to follow them when filtering.

c) Check matchings between informative and structural part of the records of the trees from Ensembl Plants. The algorithm for this check is as follows: for each group was formed data structure "information + tree" for which is checked whether the names of the genes described in SEQ - lines (the informative part) are presented in the structural part. If there are genes that are missing in one of the two parts: informative or structural, the entire line description is removed.

d) Pre-filtering of the entire dataset, where the main criterion for the purposes of comparative genomic analysis, the subject of this work is to isolate phylogenetic trees that contain data from T.aestivum and at least one of the following closely related species (T.urartu and the A. tauschii), as well as from plant model for the study of cereals - Brachypodium distachyon (Girin et al, 2013).

e) Tracking and cleaning of genes from all non-plant species.

f) Generating of lists including all available information for all genes of each species - the subject of this study.

g) Tracing the evolutionary events in phylogenetic trees and defining all possible homologous relationships between the studied species. Mapping of closest genes in the three species on bread wheat (Triticum Aestivum) sub-genomes. For all steps of the analysis (a to g) were developed Python (version 2.7) scripts.

Results and Discussion

Information used in the study from release 27 of Ensembl Plants (https://www.python.org/download/releases/2.7/) includes dataset with 56 387 gene phylogenetic trees in Newick format, containing 1 280 368 genes from 39 plants and 5 animal species. The results obtained by the applied workflow of the analysis comprised: extracting general information from phylogenetic trees, indexing the trees for tracing purposes, checking matches between informative and structural parts, cleaning genes from non-plants species using TreeNode class from E.T.E. software package (A Python Environment for Phylogenetic Tree Exploration (http://http://etetoolkit.org, Huerta-Cepas et al., 2010) - generating of lists of genes and structuring them concerning the T. aestivum sub-genomes, and tracing evolutionary events and homologous relations between species with PhyloNode of E.T.E., mapping of the closely related genes from the referent species on the three wheat sub-genomes (Figure 2).

After filtering the data from the operational phylogenetic trees, 56 387 trees were reduced to 10 056, which represents 17.8% of the total. Following the same filtration approach the number of genes from 1 280 368 was reduced to 1 002 391, which constitutes 78.3% of the total number of genes. A greater degree of keeping somehow unchanged number of genes at the expense of the changed number of phylogenetic trees can be explained by the large amount of "small" trees (over 50%), composed of about some tens of genes. Because of the few genes in their composition, these phylogenetic trees drop out massively during the step of removing single species trees.

Analyses of the evolutionary events, subject of this study were carried out with the filtered data containing the following number of genes of the studied species: T.aestivum – 94 192 genes (where 92.02% of the genes were from raw data and 95.89% from coding genes); T.urartu has 27 546 genes (90.55% of the genes – from raw data and 78.92% of the coding genes), A.tauschii 30 243 (94.03% of the genes from raw data and 97.21% - of the coding genes), B. distachyon 24 215 (94.84% of the genes from the raw data, 91.20% of the coding genes). Moreover, the clean data for the genes in each of the three sub-genomes are distributed almost equally (A - 31%, B 37.3% and D-31.7%) (Table 1).

Verification of matches between informative and structural part showed no discrepancies neither in the raw data nor after filtering. Also, after the removal of non-plant genes and the subsequent restructuring there was no loss of information, significant for the analysis such as genes of any of the studied four types or entire phylogenetic trees observed in the same process of filtering in a larger dataset (Avdjieva, 2015). Despite the differences in the data volume, this step is essential for subsequent reducing of error frequency in phylogenetic trees reconstruction.

For tracing the evolutionary events that occurred during formation of the modern T.aestivum genome, over 10 056 filtered trees was applied the algorithm Species overlap (SO), part of the E.T.E. class PhyloNode (http://etetoolkit.org/docs/2.3/tutorials/tutorial_phylogeny.htm). This algorithm was chosen to be applied on the operational dataset instead of the alternative approach Tree reconciliation, since the last requests availability of a reference species tree.
**Figure 2.** Process diagram of applied workflow.

**Table 1.** Distribution of genes in the filtered data and their share of known coding genes for the relevant species.

<table>
<thead>
<tr>
<th>Species</th>
<th>T. urartu</th>
<th>A. taushii</th>
<th>B. distachyon</th>
<th>Triticum aestivum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of genes</td>
<td>27,546</td>
<td>30,243</td>
<td>24,215</td>
<td>29,989</td>
</tr>
<tr>
<td>Proportion of coding genes</td>
<td>78.92%</td>
<td>97.21%</td>
<td>91.20%</td>
<td>31.00%</td>
</tr>
</tbody>
</table>
RESEARCH ARTICLE

SO automatically detects duplication events recognizing them for each node, commonly shared by two branches of the phylogenetic tree. For each of these evolutionary events was defined whether relations between the two edges, respectively the genes comprised by them are orthologous or paralogous. Initially all the evolutionary relationships between all the genes were determined by using the PhyloNode.get_descendant_evol_events.

In the results of this study were taken into account only the events and relationships between homologous \textit{T.aestivum} and the other three studied species, while other results were used for further analysis. For all the trees (10 056) were detected 54 807 groups of homologs, including genes of wheat, of which 35 933 groups orthologs and 18 874 groups paralogs (Table 2).

In the number of paralogs are included homeologous relations between the genes of the three wheat sub-genomes. In the homologous groups are presented all the genes from the studied four species. Their number exceeds the number of wheat genes in the data; the reason is that the algorithm to detect homologs crawls all the nodes of the tree, so that the same gene (end node) is included in as much homologous groups as are the duplication events (intermediate nodes) which remote the particular homolog from the root of the tree. The degree of homology between the genes from the root of the phylogenetic tree to it end node is increased. Therefore, for results with greatest information scores were adopted homologous groups of most distant sub-trees.

Based on these data were detected and mapped genes of the diploid predecessors on the corresponding genes in the hexaploid wheat sub-genomes. These analyses were conducted only with ortologous groups while paralogous groups were ignored because they do not carry information about the similarity in terms of functionality. For this purpose the orthologous groups were defined by the number of genes they contain, and for informative groups were considered those ones with lowest number of genes corresponding to the most distant sub-trees. In determining the closest orthologs of sub-genomes A and D were considered and mapped the genes of the two predecessors - respectively \textit{T.urartu} and \textit{A.tauschii}, while for sub-genome B, due to the lack of referent predecessor were mapped genes of \textit{B.distachyon}.

Because of the difference in the number of genes from the original data, a complete full mapping could not be performed, but for more than 50% of the genes in the sub-genomes A and D were found relevant to the closest corresponding orthologs. In the absence of the genome of the direct precursor of the B sub-genome the mapping results amounts to less than 20% of the genes as compared to the other two sub-genomes. The results of the mapping of the three genomes, although their incompleteness, were confirmed by further BLAST alignments.

Processing of the results from tracking the evolutionary events was hampered by the nature of the algorithm, which generates a large number of repetitions. This imposes further research the algorithm to be updated and improved. Also the poor mapping of the model plant \textit{B.distachyon} to homologous groups although the common origin with the other three types is intriguing. The probable reason for this is the much closer relationship between the wheat types and size of their genomes. Option for obtaining better results is the adding indispensable information from new, closely related reference types. In terms of methodology the in-house developed scripts are applied in studying of evolutionary information and assist in solving problems encountered in processing of such data.

\begin{table}[h]
\centering
\caption{Number of orthologs and paralogs.}
\begin{tabular}{lcccccc}
\hline
\textbf{Species} & \multicolumn{2}{c}{\textbf{Triticum urartu}} & \multicolumn{2}{c}{\textbf{Aegilops tauschii}} & \multicolumn{2}{c}{\textbf{Brachypodium distachyon}} \\
& orthologs & paralogs & orthologs & paralogs & orthologs & paralogs \\
\hline
\textbf{Triticum aestivum} & 27 772 & 12 869 & 28 457 & 13 432 & 22 956 & 7998 \\
\textbf{Share of all groups} & 77.29\% & 68.18\% & 79.19\% & 71.17\% & 63.89\% & 42.38\% \\
\hline
\end{tabular}
\end{table}

\textit{SPECIAL EDITION / ONLINE} Section “Cell & Molecular Biology”
Second National Youth Conference “Biological sciences for a better future”, Plovdiv, October 30-31, 2015
Conclusions

The relevance of the study concerning the comparative analysis of hexaploid wheat evolution as well as and the applied in silico methods for tracking these events can be summarized as follows:

1) Indispensable optimization of initial data normalization and processing and using data from sources other than Ensembl Plants;

2) In silico trace of the evolutionary relationships in the genome and in the sub-genomes of Triticum aestivum by enrichment the available information with new referent species.

3) Development of software solutions for comparison of the studied species to the bread wheat genome in its full structure.

4) Potential for studying the functional annotation of genes represented, as well as for predicting their functions.

Acknowledgement

This study is supported by the National Science Fund of Bulgaria within the project ”Methods for Data Analysis and Knowledge Discovery in Big Sequencing Datasets” under contract DFNI-102/7, 12th December, 2014.

References


