Genetics and Plant Physiology – 2015, Volume 5(3–4), pp. 259–266
©2015 Published by the Institute of Plant Physiology and Genetics – Bulgarian Academy of Sciences
Available online at http://www.ifrg-bg.com

PHYLOGENETIC ASSESSMENT OF THE GAIN/LOSS OF DUPLICATED GENES IN THE EVOLUTION OF GRASSES (POACEAE)

Avdjieva I.[*]

AgroBio Institute. 8, Dragan Tsankov Blvd, 1164 Sofia, Bulgaria

Received: *23 January 2015* Accepted: *13 October 2015*

Symmary: Recent paleogenomic studies have confirmed that present-day plant genomes are shaped by numerous gene/genome duplication events followed by chromosome fusions. The result from these events is the generation of copies of the coding and regulating sequences in the genome which is referred to as gene gain. The opposite process that includes removing of duplicated genes is called gene loss. Since the variation of gene copy numbers can result in certain phenotypic effects that affect biological functions it is essential to have accurate models of both gene and function gain/loss during evolution.

The present study is focused on the evaluation of gain and loss of genes in grass species (*Poaceae*) and uses phylogenetic data that is publicly available in the Ensembl Plants database. The evaluation of gain/loss of genes is done by mapping and classification of homologous gene pairs in corresponding synteny blocks. The obtained results showed that the rate of gene gain/loss varied between different grass species and was influenced by the genome size and structure.

Citation: Avdjieva I., 2015. Phylogenetic assessment of the gain/loss of duplicated genes in the evolution of grassess (Poaceae). *Genetics and Plant Physiology*, 5(3–4): 259–266.

Keywords: Evolution; gene gain; gene loss; grasses; synteny; homology.

Abbreviations: CNV – Copy Number Variation; CoGe – COmparative GEnomics; E. T. E. – (python) Environment for (phylogenetic) Tree Exploration; EBI – European Bioinformatics Institute; EBML - European Molecular Biology Laboratory; iTOL – interactive Tree Of Life; SSD - Small Scale Duplications; WGD - Whole Genome Duplication.

NTRODUCTION

It has been several decades since Susumu Ohno proposed that modern diploids have originated from paleopolyploids by means of ancestral chromosome fusions and sequence divergence between chromosomes (Ohno, 1970). Paleogenomic analyses in plants confirmed and refined Ohno's conclusions by discovering that several rounds of Whole Genome Duplications (WGDs), Small Scale Duplications (SSDs), Copy Number Variations (CNVs) and chromosome fusion events have shaped the chromosome number observed in modern plants (Pont et al., 2011; Abrouk et al., 2010). Polyploidy followed by diploidization is considered a major mechanism that has formed complex

[*]Corresponding author: i.y.stoycheva@gmail.com

regulatory networks during the evolution of plants, thus the role of duplications needs to be assessed.

WGDs and SSDs in plants result in multiple copies of the coding and regulating sequences in the genome which is referred to as gene gain. Duplicated genes that are retained during evolution are called persistent and are prone to at least partial functional divergence such as: 1) unexpected/ functionless paralog; 2) partitioned function; 3) novel function. The last two events are considered important sources of evolutionary innovation in organisms (Doyle et al., 2008). The opposite process – gene loss, involves removing of duplicated genes by fractionation. This process refers to mutations leading to the loss of redundant function by any of the following: randomization by substitution of neutral base pairs, deletion, insertion, copy over by simple sequence repeats (SSRs), and similar processes. (Langham et al., 2004). The loss of genes is biased in multiple ways – in the retention of duplicated copies and in the loss between duplicated regions. This is a general property of eukaryotic gene and genome duplications (Sankoff et al., 2010) and may represent a useful mark for ancestral subgenome reconstructions (Schnable et al., 2012).

Analyses of both syntenic and homologous relationships between genes can give insight of duplicated gene loss/ gain patterns during the course of evolution. In modern genomics the term ‘synteny’ means conservation of gene blocks within two sets of chromosomes from different species regardless of whether they are genetically linked (Renwick, 1972; Passarge et al., 1999). This is also referred to as shared synteny and is one of the most reliable criteria for establishing orthology between genomic regions in different species. Rearrangements to the genome may result in the loss or gain of synteny between loci (Moreno-Hagelsieb et al., 2001). Such patterns can be used to explore phylogenetic relationships between species and to infer the genome organization of extinct ancestors.

The main goal of this study was to propose a method that analyzes the gene gain/loss patterns based on *in silico* analysis of syntenic regions (position) and phylogenetic trees (relationship) between grasses. It required several stages: 1) obtaining and processing homologous data; 2) obtaining and processing synteny data; 3) mapping both datasets; 4) analysis of the matching data.

MATERIALS AND METHODS

The source data consists of phylogenetic trees from the public database Ensembl Plants (Kersey et al., 2013; ftp://ftp.ensemblgenomes.org/pub/plants/release-19/emf/compara/homologies/) which is developed and supported by the European Bioinformatics Institute, part of the European Molecular Biology Laboratory (EMBL-EBI). The database Phytozome (Goodstein et al., 2012; http://phytozome.jgi.doe.gov/pz/portal.html) was used as a cross-reference for general information about the genomes. The data originally contained 884014 genes from 23 plant and 5 animal species, organized in 43771 phylogenetic trees. A short description and a multiple sequence alignment of the genes in each tree were also available.

The data was processed according to several criteria: 1) Trees containing genes from only one species were considered not informative and removed. 2) Genes that do not belong to plants were traced and removed from the trees. 3) Genes located in scaffolds and plastids were also removed. Exceptions were made only if a scaffold was listed in both Ensembl and Phytozome as containing high percentage of genome sequencing data. After each change in the gene content of a tree it was first reconstructed and then a verification of the first criterion was carried out.

The dataset was processed by custom scripts written in Python, version 2.7. The manipulation and reconstruction of the trees involved a Python-based programming toolkit called A Python Environment for (phylogenetic) Tree Exploration (E. T. E.) (Huerta-Cepas et al., 2010; http://etetoolkit.org) that also required predominantly custom scripts. The visualization of phylogenetic trees was done by the on-line tool iTOL (Letunic and Bork, 2006; http://itol.embl.de) which is also part of the EMBL.

The syntenic relationships required for the gene dynamics analysis were obtained in the following pipeline: 1) generation of synteny blocks from the whole genome public data; 2) mapping with the processed phylogenetic trees; 3) studying the dynamics of homologous genes within corresponding synteny blocks. The data was generated with SynMap, part of the Comparative Genomics (CoGe) toolset (Lyons and Freeling 2008; https://genomevolution.org/coge/SynMap.pl). SynMap is an on-line tool which generates a synteny map between the chromosomes of two organisms and identifies syntenic regions. First, every coding sequence is compared between the taxa using BLASTn (Altschul et al., 1997) in order to identify homologous gene pairs. Those were processed by DAGChainer (Haas et al., 2004) to find collinear sets of genes shared between the taxa. The results from SynMap were presented as combined datasets mapped according to their relative genomic position where homologous gene pairs are plotted in grey and syntenic gene pairs are plotted in color. For the purposes of this study only syntenic pairs were taken into account. Then, using Python scripts, a consistency check between the datasets was performed and the analysis was carried out only with the matching genes. The dynamics of syntenic blocks during the course of evolution and the presence/absence of orthologous genes were traced using the species tree from iTOL as a guide.

RESULTS AND DISCUSSION

The results from the processing of the phylogenetic trees are summarized in Figure 1. During the process around 12% of the genes but more than 50% of the trees were removed. This bias is explained by: 1) large amount of "small" trees containing less than five genes, 2) large amount of single-species trees and 3) trees lost during reconstruction (around 10% of all trees). A possible explanation of the last issue is that in such cases the removal of root genes leads to impossibility to reconstruct the tree.

The "clean" dataset contained nine grass species - *Brachypodium distachyon, Hordeum vulgare, Oryza brachyantha, Oryza glaberrima, Oryza indica, Oryza sativa, Setaria italica, Sorghum bicolor, Zea mays*.

The genomes for the synteny analysis were predefined by the CoGe which

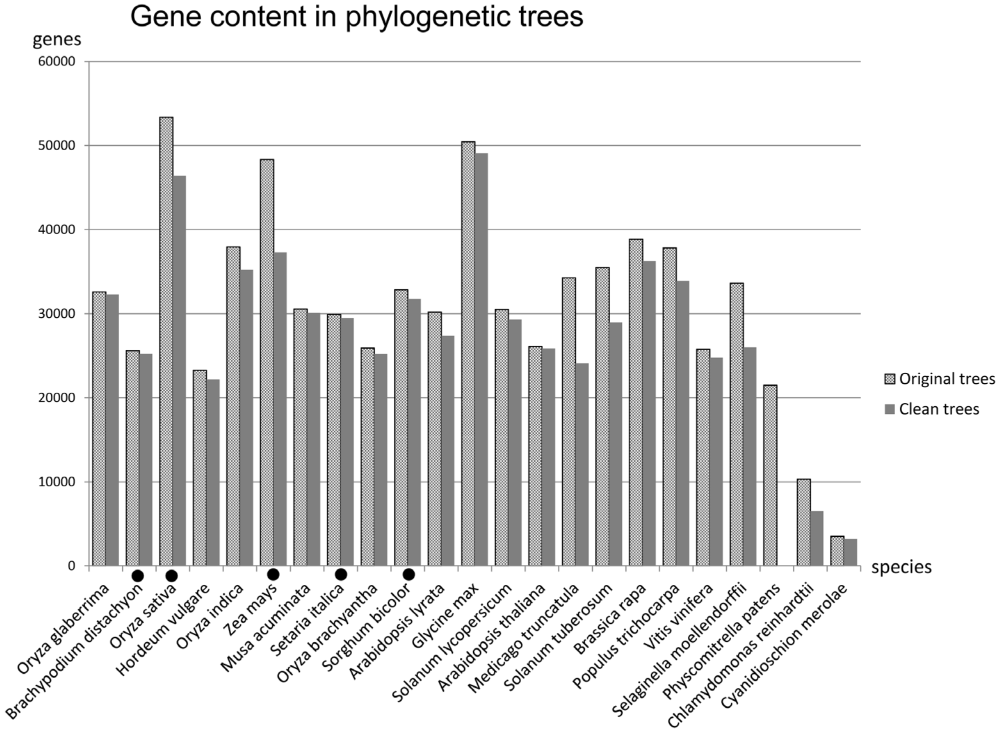

Figure 1. Comparative chart of the gene content (genes per species) before and after the procession of the phylogenetic dataset. The species included in the current research are marked with black dots.

caused several problems: 1) not all grasses from the phylogenetic dataset were present (*Oryza glaberrima*); 2) some genomes did not match the genome database version or chromosome content (*Hordeum vulgare, Oryza brachyantha, Oryza indica*). The analysis the genomes of the four *Oryza* species showed high similarity and *O. sativa* (rice) was selected to represent them but *Hordeum vulgare* (barley) had to be excluded. The research was continued with the following species: *B. distachyon* (purple false brome)*, O. sativa* (rice)*, S. italica* (foxtail millet)*, S. bicolor* (sorghum) and *Z. mays* (maize). All five species were plotted against each other and the output from the genome-to-genome synteny mapping showed relatively uniform results with an average count of 750 synteny blocks. Only in the *Z. mays* pairs an average of 1100 synteny blocks was detected. This bias is probably due to maize's most resent WGD that results in higher transcript count. A similar pattern is seen in the number of genes included in the synteny blocks (an average of 23000 pairs). Here the highest gene count (28000) is in the maize/sorghum pair.

The comparison between the gene content of the synteny blocks and the phylogenetic trees revealed that almost 90% of the synteny genes are present in the trees. The genes found in both datasets were analyzed following two parallel directions: dynamics of synteny blocks during the course of evolution according to the species tree of grasses (see Fig. 2) and dynamics of the genes within the

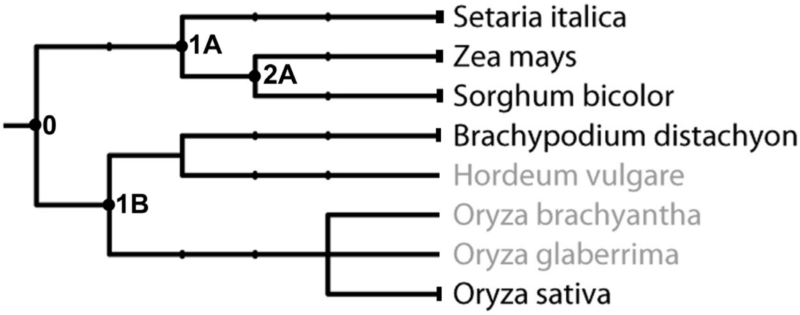

Figure 2. Species tree of grasses illustrating the speciation events (black dots) during evolution. The two groups (A and B) and two subgroups (1 and 2) of genes allow tracing the conservation of synteny blocks. The species marked in grey were excluded from the research due to insufficient or contradicting information between the data in the trees and the genomes used by SynMap.

blocks according to tree topologies.

In the first case the synteny blocks were regrouped according to the speciation events and each two groups were compared to see if a certain block was conserved, merged with another (gain of syntgeny) or split (loss of synteny). The rate of split/merged synteny blocks was biased according to their size but a distinct pattern could not be established.

As an example a synteny map between the genomes of *Oryza sativa* and *Brachypodium distachyon* is shown in Fig. 3. It shows the synteny blocks as scattered dots which form a diagonal line where large areas of the corresponding chromosomes show synteny. Also, a more complex image of all genomes plotted against *Z. mays* was generated. Maize was chosen as a reference for being both the largest and the "youngest" genome in the current dataset.

The dynamics of gene gain/loss was assessed in all trees and subtrees (two or more nodes (children) sharing an ancestor (parent)) containing at least two grass species sharing a synteny block. The presence of all five species within a subtree and synteny block indicated gene retention. Other cases (around 10 % of the genes) showed more genes per species than expected in the subtree and were considered gains. In subtrees where at least one of five species were missing or contained fewer copies than the others were considered losses. Either way, the results showed that the rate of gene gain/loss was biased between different grass species and was influenced by the genome size and structure.

There are two major outcomes from this study and both of them provide a basis for further research but also require additional attention and, perhaps, an alternative approach. The method itself is suitable for processing phylogeny data from other sources and organisms after normalization. During the course of this study several flaws in the dataset were revealed and are being solved in the present. First, the processing method of the source data can benefit from indexation in order to better understand and trace individual trees. A small part of the

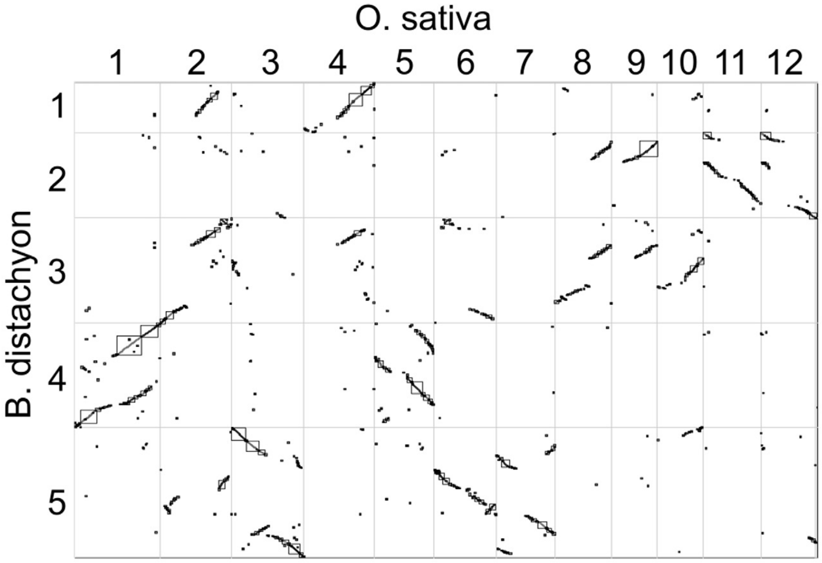

Figure 3. Synteny map of the chromosomes of *Oryza sativa* and *Brachypodium distachyon*. The numbers of the rows indicate the chromosomes of *Brachypodium distachyon* and the numbers of the columns indicate the corresponding chromosomes of the *Oryza sativa*. The cells represent synteny dot-plot between each two chromosomes. Syntenic regions are represented by black dots and outlined by rectangles.

data contained mismatched information on different levels that are caused by imperfections of the database and need to be reviewed separately. The overall speed of the process can also be optimized.

As mentioned before, the synteny mapping process was restricted from the predefined choice of genomes to work with. Also, the lack of verification source may lead to misinterpretation of the results. These issues bring up the need of a custom tool with more functionality that would allow custom gene datasets to be directly mapped for (partial) synteny assessment either against predefined whole genomes or against other datasets. The combined assessment of different sequencing versions of the genomes is also an opportunity.

CONCLUSIONS

Variations in the genome, regulome and phenotypic characteristic in grass species, as well as their major involvement in agriculture makes them a suitable model to study evolutionary events and train *in silico* methods. The “clean” phylogenetic trees, obtained from the processing method described in this study, are ready to be used as an input for various topology-based researches of plant evolution including comparative evolution of various major phenotype characteristics. The combined synteny and homology approach sheds some light onto the dynamics of duplicated gene content during evolution but also asks for more complex analyses to be conducted.

ACKNOWLEDGEMENTS

Dr Dimitar Vassilev from AgroBio Institute, Dr Jerome Salse from INRA, France and acad. Atanas Atanassov from Joint Genomic Centre are sincerely acknowledged for their support.

REFERENCES

Abrouk M, F Murat, C Pont, J Messing, S Jackson, T Faraut, E Tannier, C Plomion, R Cooke, C Feuillet, J Salse, 2010. Palaeogenomics of plants: synteny-based modelling of extinct ancestors. Trends in Plant Science, 15: 479–487.

Altschul S F, L T Madden, A A Schäffer, J Zhang, Z Zhang, W Miller, D J Lipman, 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Doyle J J, L E Flagel, A H Paterson, R A Rapp, D E Soltis, P S Soltis, J F Wendel, 2008. Evolutionary genetics of genome merging and doubling in plants. Annu Rev Genet. 42: 443–461.

Goodstein D, S Shu, R Howson, R Neupane, R D Hayes, J Fazo, T Mitros, W Dirks, U Hellsten, N Putnam, D S Rokhsar, 2012. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 40 (D1): D1178–D1186.

Haas B J, A L Delcher, J R Wortman, S L Salzberg, 2004. DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics 20(18): 3643–3646.

Huerta-Cepas J, J Dopazo, T Gabaldón, 2010. ETE: a python Environment for Tree Exploration. BMC Bioinformatics, 11: 24.

Kersey P J, J Allen, M Christensen, P Davis, L J Falin, C Grabmueller, D Seth, T Hughes, J Humphrey, A Kerhornou, J Khobova, N Langridge, M McDowall, U Maheswari, G Maslen, M Nuhn, C K Ong, M Paulini, H Pedro, I Toneva, M A Tuli, B Walts, G Williams, D Wilson, K Youens-Clark, M K Monaco, J Stein, X Wei, D Ware, D M Bolser, K L Howe, E Kulesha, D Lawson, D M Staines, 2014. Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic acids research, 42 (D1): D546–D552.

Langham R J, J Walsh, M Dunn, C Ko, S A Goff, M Freeling, 2004. Genomic duplication, fractionation and the origin of regulatory novelty. Genetics 166: 935–945.

Letunic I and P Bork, 2006. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23(1):127–128.

Lyons E and M Freeling, 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. The Plant Journal 53:661–673.

Lyons E, B Pedersen, J Kane, M Alam, R Ming, H Tang, X Wang, J Bowers, A Paterson, D Lisch, 2008. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids. Plant Phys 148: 1772–1781.

Moreno-Hagelsieb G, V Treviño, E Pérez-Rueda, T F Smith, J Collado-Vides, 2001. Transcription unit conservation in the three domains of life: a perspective from Escherichia coli. Trends in Genetics 17 (4): 175–177.

Ohno S, 1970. Evolution by Gene Duplication. New York: Springer-Verlag, pp. 160.

Passarge E, B Horsthemke1, R A Farber, 1999. Incorrect use of the term synteny. Nat Genet 23: 387.

Pont C, F Murat, C Confolent, S Balzergue, J Salse, 2011. RNA-seq in grain unveils fate of neo- and paleopolyploidization events in bread wheat (*Triticum aestivum* L.). Genome Biology 12: R119.

Python Software Foundation. Python Language Reference, version 2.7.

Renwick J H, 1972. Proceedings of the Fourth International Congress of Human Genetics, 6-11 September 1971. Excerpta Medica, Amsterdam, 443–444.

Sankoff D, C Zheng, Q Zhu, 2010. The collapse of gene complement following whole genome duplication. BMC Genomics 11: 313.

Schnable J C, M Freeling, E Lyons, 2012. Genome-wide analysis of syntenic gene deletion in the grasses, Genome Biology and Evolution 01 4(3):265–277.