# Computational analysis of HD-Zip transcription factor gene family from *Solanum tuberosum* L.

Ashutosh Mukherjee
Department of Botany, Vivekananda College, 269, Diamond Harbour Road, Thakurpukur, Kolkata - 700063, West Bengal, India.

***Abstract :*** In the plant kingdom, a unique family of transcription factors are found called HD-Zip (homeodomain leucine zipper) transcription factors. They show the combination of a homeodomain with a leucine zipper acting as a dimerization motif. They have roles in plant growth including organ and vascular development or meristem maintenance as well as response to environmental stresses. In several crop plant species, these proteins showed abiotic stress response. These findings imply HD-Zip genes as high priority candidates for improvement of crop species. *Solanum tuberosum* L. (potato), a member of Solanaceae, is the world's most important non-grain food crop. The publicly available genome sequence of this species provides a platform for genetic improvement of this crop. In this study, computational analysis of the HD-Zip gene family of this plant has been performed. There are seventy seven members of this gene family in *S. tuberosum* encoded by fifty eight genes. Phylogenetic analysis divided the proteins into four different groups. Study of Gene Ontology revealed that some of these genes participate in many crucial biological processes in *S. tuberosum* including response to salt stress, response to blue light and xylem and phloem pattern formation to name a few. Study of some genic features revealed GC content and length variation among these genes. This study provides a platform for agronomic improvement of *S. tuberosum* and other crops by several biotechnological approaches regarding HD-Zip transcription factors.

***IndexTerms*** **– Clade, GC content, Gene Ontology, homeodomain, phylogenetic tree**

---

## I. INTRODUCTION

homeodomain (HD) is a conserved 60-amino acid motif present in transcription factors found in all the eukaryotic organisms and this 60-amino acid sequence folds into a characteristic three-helix structure that is able to interact specifically with DNA (Ariel et al., 2007). Most HDs are able to bind DNA as monomers with high affinity, through interactions made by helix III (also called as recognition helix) and a disordered N-terminal arm located beyond helix I (Ariel et al., 2007). There is high degree of conservation of this type of domain among diverse proteins found in species of different kingdoms which indicates that the typical structure of these proteins is crucial to maintain the HD functionality and also this domain plays vital role (Moens and Selleri, 2006). The members of the plant homeodomain superfamily differ in the sequence encoding the HD, their size, HD location, association with other domains and in their genes structures (Ariel et al., 2007). Based on these distinguishing features, plant HD-containing proteins can be classified into six families namely HD-Zip (homeodomain associated to a leucine zipper), PHD finger (plant homeodomain associated to a finger domain), Bell (named after the distinctive Bell domain), ZF-HD (zinc finger associated to a homeodomain), WOX (Wuschel related homeobox) and KNOX (Knotted related homeobox) (Ariel et al., 2007).

HD-Zip (Homeodomain leucine zipper) transcription factors have been found exclusively in the plant kingdom (Ariel et al., 2007; Mukherjee et al., 2009), except in the charophycean algae (Zalewski et al., 2013). The characteristic feature of the HD-Zip gene family is the association of homeodomain (HD) and the leucine zipper (LZ) motif in a single protein while in other kingdoms; they are present as domains of distinct proteins (Belamkar et al., 2014). HD-Zip proteins bind to DNA as dimers and leucine zipper (LZ) acts as the dimerization motif while the HD is responsible for the specific binding to DNA (Ariel et al., 2007). The HD-Zip transcription factors can be subdivided into four subfamilies: HD-Zip I to IV (Belamkar et al., 2014), by the following four distinguishing features: (i) conservation of the HD-Zip domain that determine DNA-binding specificities, (ii) genes structures, (iii) additional conserved motifs and (iv) functions (Ariel et al., 2007). HD-Zip genes have been found to be involved in meristem regulation, photomorphogenesis, root development and several abiotic stress responses (Ariel et al., 2007; Harris et al., 2011). The HD-Zip I genes have been investigated for their roles in water deficit and salt stress responses (Belamkar et al., 2014). The HD-Zip superfamily has been studied in several species including *Arabidopsis thaliana* (Henriksson et al., 2005; Ciarbelli et al., 2008), *Oryza sativa* (Agalou et al., 2008), *Zea mays* (Zhao et al., 2011) and *Populus trichocarpa* (Hu et al., 2012). This shows the importance of proper characterization of this transcription factors in plant biology.

*Solanum tuberosum* L. (potato) is a member of the family Solanaceae and this plant is the world's most important non-grain food crop (The Potato Genome Sequencing Consortium, 2011). *S. tuberosum* occupies a wide eco-geographical range and its importance is growing rapidly, especially within the developing world (The Potato Genome Sequencing Consortium, 2011). Worldwide, the tubers of this plant are an important dietary source of starch, protein, antioxidants and vitamins (Burlingame et al., 2009). There are significant barrier to potato improvement using classical breeding approaches as this plant is highly heterozygous, suffer acute inbreeding depression and are susceptible to many devastating pests and pathogens (The Potato Genome Sequencing Consortium, 2011). To overcome these limitations, knowledge of the structural and functional aspects of different gene families of potato is important. Considering the significance of HD-Zip transcription factors in different aspects of plant development and environmental interactions, the present investigation has been performed to characterize the HD-Zip gene families of *S. tuberosum* using publicly available genomic data of the plant. The objective of the present investigation was to study a) the phylogenetic

relationships among the HD-Zip proteins, b) different genic features of these proteins and c) functional characterization of these transcription factors.

## II. RESEARCH METHODOLOGY

Protein sequences of the HD-Zip transcription factors of *Solanum tuberosum* L. have been obtained from Plant Transcription Factor Database v4.0 (Jin et al., 2017) and corresponding gene names have been obtained from Phytozome (Goodstein et al., 2012). Gene start and end positions (base pairs) as well as % GC content (percentage of Guanine+Cytosine nucleotides) of these genes have been retrieved using the BioMart tool (Kinsella et al., 2011) of Ensembl Plants database (Kersey et al., 2018). Phylogenetic analysis using the protein sequences was performed with MEGA version 5 (Tamura et al., 2011) by Neighbor-Joining method (Saitou and Nei, 1987) after performing a multiple sequence alignment with ClustalW (Thompson et al., 1994). The resulting phylogenetic tree was visualized with FigTree downloaded from http://beast.bio.ed.ac.uk/FigTree. For functional characterization of the HD-Zip proteins, GO *i.e.* Gene Ontology (The Gene Ontology Consortium, 2012) terms associated with these proteins have been retrieved using the same Biomart tool of Ensembl Plants database. All three categories of GO domains namely biological processes, molecular functions and cellular components have been analyzed. Wilcoxon Rank Sum test (Wilcoxon, 1945) was used for pair-wise comparisons of different variables since the values were not normally distributed in the datasets investigated in the present study. All statistical analyses were performed with R package (R Core Team, 2018).

## III. RESULTS AND DISCUSSION

### Phylogenetic relationships among the HD-Zip proteins of *Solanum tuberosum* L.

The phylogenetic tree of the HD-Zip proteins of *S. tuberosum* is shown in figure 1. All the 77 HD-Zip proteins have been grouped into four clades (shown in green, blue, red and purple, respectively in figure 1).

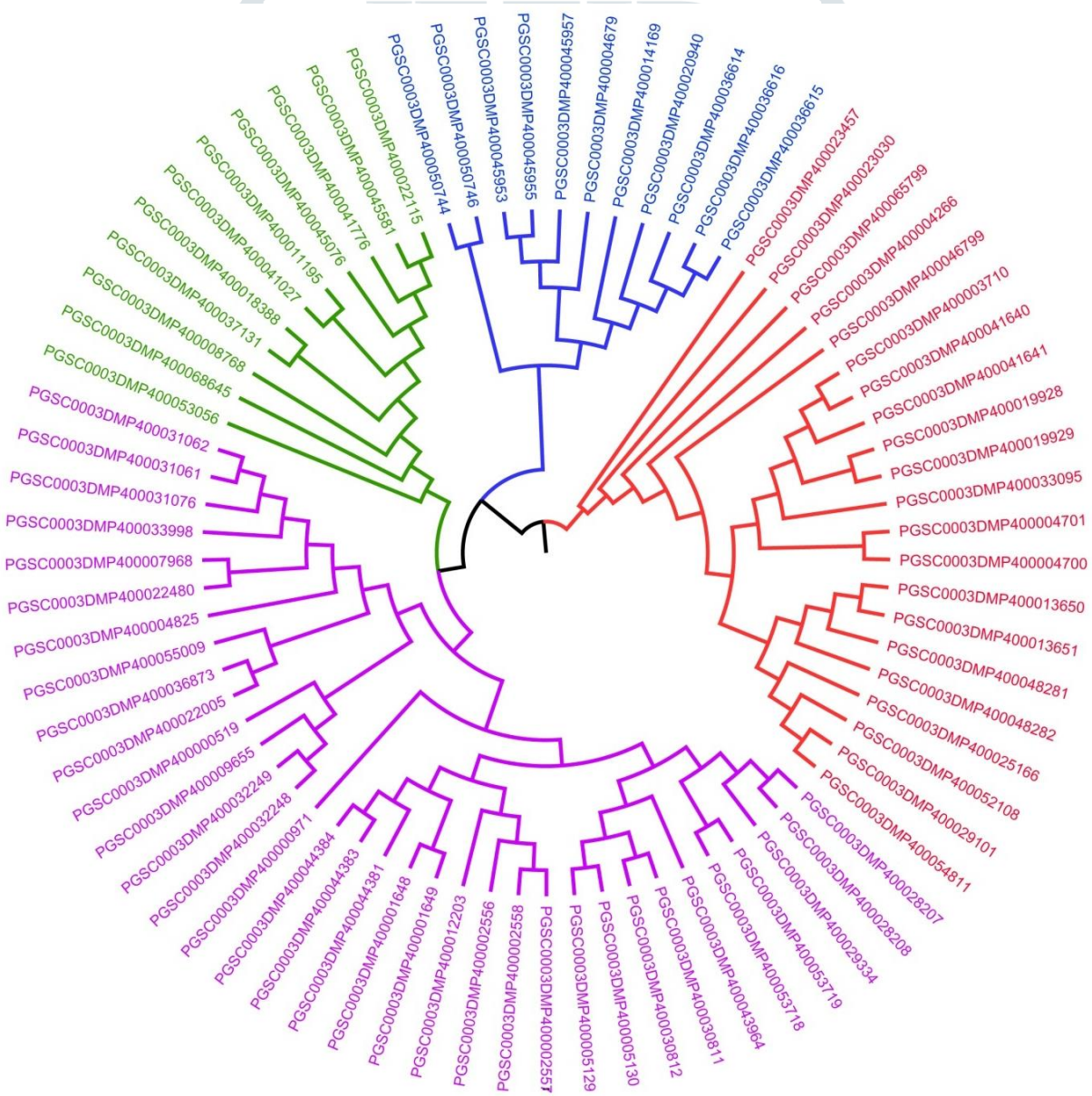


**Figure 1:** Phylogenetic tree of the 77 HD-Zip transcription factors based on protein sequences. The HD-Zip proteins have been grouped into four clades (shown in green, blue, red and purple).

The first group (green) comprised of 11 members (PGSC0003DMP400022115, PGSC0003DMP400045581, PGSC0003DMP400041776, PGSC0003DMP400045076, PGSC0003DMP400011195, PGSC0003DMP400041027, PGSC0003DMP400018388, PGSC0003DMP400037131, PGSC0003DMP400008768, PGSC0003DMP400068645 and PGSC0003DMP400053056). The second group (blue) also comprised of 11 members (PGSC0003DMP400050744, PGSC0003DMP400050746, PGSC0003DMP400045953, PGSC0003DMP400045955, PGSC0003DMP400045957, PGSC0003DMP400004679, PGSC0003DMP400014169, PGSC0003DMP400020940, PGSC0003DMP400036614, PGSC0003DMP400036616 and PGSC0003DMP400036615). The third group (red) comprised of 21 members (PGSC0003DMP400023457, PGSC0003DMP400023030, PGSC0003DMP400065799, PGSC0003DMP400004266, PGSC0003DMP400046799, PGSC0003DMP400003710, PGSC0003DMP400041640, PGSC0003DMP400041641, PGSC0003DMP400019928, PGSC0003DMP400019929, PGSC0003DMP400033095, PGSC0003DMP400004701, PGSC0003DMP400004700, PGSC0003DMP400013650, PGSC0003DMP400013651, PGSC0003DMP400048281, PGSC0003DMP400048282, PGSC0003DMP400025166, PGSC0003DMP400052108, PGSC0003DMP400029101 and PGSC0003DMP400054811) and the fourth (the largest) group (purple) comprised of 34 members (PGSC0003DMP400028207, PGSC0003DMP400028208, PGSC0003DMP400029334, PGSC0003DMP400053719, PGSC0003DMP400053718, PGSC0003DMP400043964, PGSC0003DMP400030811, PGSC0003DMP400030812, PGSC0003DMP400005130, PGSC0003DMP400005129, PGSC0003DMP400002557, PGSC0003DMP400002558, PGSC0003DMP400002556, PGSC0003DMP400012203, PGSC0003DMP400001649, PGSC0003DMP400001648, PGSC0003DMP400044381, PGSC0003DMP400044383, PGSC0003DMP400044384, PGSC0003DMP400000971, PGSC0003DMP400032248, PGSC0003DMP400032249, PGSC0003DMP400009655, PGSC0003DMP400000519, PGSC0003DMP400022005, PGSC0003DMP400036873, PGSC0003DMP400055009, PGSC0003DMP400004825, PGSC0003DMP400022480, PGSC0003DMP400007968, PGSC0003DMP400033998, PGSC0003DMP400031076, PGSC0003DMP400031061 and PGSC0003DMP400031062). The members of group 4 were again divided into two subgroups. The smaller subgroup comprised of 14 members and the larger subgroup comprised of 20 members. This phylogenetic grouping supports the view that HD-Zip transcription factors can be subdivided into four subfamilies: HD-Zip I to IV (Ariel et al., 2007; Belamkar et al., 2014).

**Genic properties of the HD-Zip transcription factors of *S. tuberosum***

It has been observed that the 77 HD-Zip proteins have been encoded by only 58 genes. There are 4 genes (PGSC0003DMG400001417, PGSC0003DMG400021125, PGSC0003DMG400025614 and PGSC0003DMG400026460) which code for three proteins each and there are 11 genes (PGSC0003DMG400000863, PGSC0003DMG400002627, PGSC0003DMG400002835, PGSC0003DMG400007733, PGSC0003DMG400011256, PGSC0003DMG400016148, PGSC0003DMG400017640, PGSC0003DMG400018509, PGSC0003DMG400024068, PGSC0003DMG400027770 and PGSC0003DMG400030837) which code for 2 proteins each. Notably, none of these genes code for group 1 HD-Zip proteins. Different genic properties of the members of the four groups of HD-Zip transcription factors of *S. tuberosum* are shown in table 1.

Table 1: Different genic properties of the members of the four groups of HD-Zip transcription factors of *S. tuberosum.*

| **Parameters (Mean±SE)** | **Group 1** | **Group 2** | **Group 3** | **Group 4** |
|---|---|---|---|---|
| Protein length (amino acids) | 280.18±15.5 | 774.82±34.17 | 232.88±10.1 | 438.34±28.69 |
| Gene length (base pairs) | 2245.18±245.60 | 6932.29±620.67 | 5559.19±424.10 | 2564.83±368.64 |
| GC content of Gene | 33.35±0.97 | 37.49±0.19 | 33.50±0.71 | 33.27±0.56 |

From table 1, it is clear that the four groups differ considerably regarding gene and protein length as well as GC content of genes. It is notable that regarding protein length, significant differences (Wilcoxon Rank Sum test, $p < 0.05$) were found between the four groups. However, regarding gene length, significant differences have been found between group 1 and 2, group 1 and 3, group 2 and 4 as well as group 3 and 4 (Wilcoxon Rank Sum test, $p < 0.01$). Regarding GC content of genes, group 2 genes showed significantly higher GC content compared to other groups (Wilcoxon Rank Sum test, $p < 0.01$). Results of protein as well as gene length variation indicate that gene structural features should be further studied elaborately in this class of genes of *S. tuberosum* for better understanding of their exon intron structure and splice variants. It has been proposed that GC content diversity has biological and evolutionary significance (Singh et al., 2016). In the present study, group 2 HD-Zip genes showed significantly higher GC content than the other three groups indicating different evolutionary history of this group of genes compared to the other groups. Further study is required to reveal the actual cause of these GC content variation.

**Functional annotations of the HD-Zip transcription factors of *S. tuberosum***

Analysis of the GO terms revealed that all the members of the HD-Zip family are expressed in nucleus as revealed by the GO domain ‘cellular component’. However, the GO domain ‘molecular function’ showed six GO term names namely ‘DNA binding’, ‘DNA-binding transcription factor activity’, ‘lipid binding’, ‘protein homodimerization activity’, ‘sequence-specific DNA binding’ and ‘transcription regulatory region DNA binding’. Regarding GO domain ‘biological process’, one GO term name ‘regulation of transcription, DNA-templated’ was predominant. Notably, several other terms have been found. These are 'cell differentiation', 'cell proliferation', 'cotyledon morphogenesis', 'determination of bilateral symmetry', 'leaf morphogenesis', 'maintenance of floral organ identity', 'meristem initiation', 'negative regulation of transcription, DNA-templated', 'negative regulation of translation', 'polarity specification of adaxial/abaxial axis', 'positive regulation of cell differentiation', 'positive regulation of cell proliferation', 'positive regulation of transcription, DNA-templated', 'primary root development', 'primary shoot apical meristem specification', 'procambium histogenesis', 'radial pattern formation', 'red or far-red light signaling pathway', 'response to auxin', 'response to blue light', 'response to cytokinin', 'response to far red light', 'response to salt stress', 'response to sucrose', 'root development', 'shade avoidance', 'shoot system morphogenesis', 'unidimensional cell growth', 'xylem and phloem

pattern formation' and 'xylem development'. These results showed that HD-Zip transcription factors of *S. tuberosum* play very important roles in this plant species. Previously, this class of transcription factors was found to play many roles in different plant species. For example, in *Craterostigma plantagineum*, CpHB-4 and CpHB-5 (class I HD-Zip genes) were down-regulated in dehydration response in leaves and roots while CpHB-6 and CpHB-7 (also class I HD-Zip genes) were found to be up-regulated (Deng et al., 2002). Turchi et al. (2013) showed that HD-Zip II transcription factors of *Arabidopsis thaliana* control apical embryo development and meristem function. Vernoud et al. (2009) showed that in maize, the HD-ZIP IV transcription factor OCL4 is necessary for trichome patterning and development of anther. The present study showed that HD-Zip transcription factors play various crucial roles in *S. tuberosum*.

In conclusion, the present study showed that the HD-Zip transcription factors of *Solanum tuberosum* can be divided into four groups based on protein sequences. Some of the genes code more than one proteins of HD-Zip family of transcription factors. The genes vary in their length and GC content. These proteins perform several important biological roles in *S. tuberosum*. This study will serve as a foundation for further studies with HD-Zip transcription factors in potato as well as other crop species which will be ultimately helpful in agronomic improvement of these crops.

## IV. Acknowledgment

The facility situated at the Department of Botany, Vivekananda College, Thakurpukur, Kolkata-700063 is gratefully acknowledged.

## References

**[1]** Agalou, A., Purwantomo, S., Övernäs, E., Johannesson, H., Zhu, X., Estiati, A., de Kam, R. J., Engström, P., Slamet-Loedin, I. H., Zhu, Z., Wang, M., Xiong, L., Meijer, A. H. and Ouwerkerk, P. B. F. 2008. A genome-wide survey of HD-Zip genes in rice and analysis of drought-responsive family members. Plant Molecular Biology, 66(1-2): 87-103.

**[2]** Ariel, F. D., Manavella, P. A., Dezar, C. A. and Chan, R. L. 2007. The true story of the HD-Zip family. Trends in Plant Science, 12(9): 419-426.

**[3]** Belamkar, V., Weeks, N. T, Bharti, A. K., Farmer, A. D., Graham, M. A., Cannon, S. B. 2014. Comprehensive characterization and RNA-Seq profiling of the HD-Zip transcription factor family in soybean (*Glycine max*) during dehydration and salt stress. BMC Genomics, 15: 950.

**[4]** Burlingame, B., Mouillé, B. and Charrondière, R. 2009. Nutrients, bioactive non-nutrients and anti-nutrients in potatoes. Journal of Food Composition and Analysis, 22(6): 494-502.

**[5]** Ciarbelli, A. R., Ciolfi, A., Salvucci, S., Ruzza, V., Possenti, M., Carabelli, M., Fruscalzo, A., Sessa, G., Morelli, G. and Ruberti, I. 2008. The Arabidopsis Homeodomain-leucine zipper II gene family: diversity and redundancy. Plant Molecular Biology, 68(4-5): 465-478.

**[6]** Deng, X., Phillips, J., Meijer, A., Salamini, F. and Bartels, D. 2002. Characterization of five novel dehydration-responsive homeodomain leucine zipper genes from the resurrection plant *Craterostigma plantagineum*. Plant Molecular Biology, 49(6): 601-610.

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

**[8]** Harris, J. C., Hrmova, M., Lopato, S. and Langridge, P. 2011. Modulation of plant growth by HD-Zip class I and II transcription factors in response to environmental stimuli. New Phytologist, 190(4): 823-837.

**[9]** Henriksson, E., Olsson, A. S. B., Johannesson, H., Johansson, H., Hanson, J., Engström, P., Söderman, E. 2005. Homeodomain leucine zipper class I genes in Arabidopsis. Expression patterns and phylogenetic relationships. Plant Physiology, 139(1):509-518.

**[10]** Hu, R., Chi, X., Chai, G., Kong, Y., He, G., Wang, X., Shi, D., Zhang, D. and Zhou, G. 2012. Genome-wide identification, evolutionary expansion, and expression profile of Homeodomain-Leucine zipper gene family in poplar (*Populus trichocarpa*). PLoS One, 7(2): e31149.

**[11]** Jin, J., Tian, F., Yang, D-C., Meng, Y-Q., Kong, L., Luo, J. and Gao, G. 2017. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Research, 45(D1): D1040-D1045.

**[12]** Kersey, P. J., Allen, J. E., Allot, A., Barba, M., Boddu S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Grabmueller, C., Kumar, N., Liu, Z., Maurel, T., Moore, B., McDowall, M. D., Maheswari, U., Naamati, G., Newman, V., Ong, C. K., Paulini, M., Pedro, H., Perry, E., Russell, M., Sparrow, H., Tapanari, E., Taylor, K., Vullo, A., Williams, G., Zadissia, A., Olson, A., Stein, J., Wei, S., Tello-Ruiz, M., Ware D., Luciani, A., Potter, S., Finn, R. D., Urban, M., Hammond-Kosack, K. E., Bolser, D. M., De Silva, N., Howe, K. L., Langridge, N., Maslen, G., Staines, D. M. and Yates A. 2018. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Research, 46(D1): D802-D808.

**[13]** Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P. and Flicek, P. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database, 2011: bar030.

**[14]** Moens, C. B. and Selleri, L. 2006. Hox cofactors in vertebrate development. Developmental Biology, 291(2): 193–206.

**[15]** Mukherjee, K., Brocchieri, L. and Bürglin, T. R. 2009. A comprehensive classification and evolutionary analysis of plant homeobox genes. Molecular Biology and Evolution, 26(12): 2775-2794.

**[16]** R Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org.

**[17]** Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4(4):406-425.
**[18]** Singh, R., Ming, R. And Yu, Q. 2016. Comparative analysis of GC content variations in plant genomes. Tropical Plant Biology, 9(3): 136–149.
**[19]** Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution, 28(10): 2731-2739.
**[20]** The Gene Ontology Consortium. 2012. The Gene Ontology: enhancements for 2011. Nucleic Acids Research, 40(D1): D559-D564.
**[21]** The Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. Nature, 475(7355): 189-195.
**[22]** Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22(22): 4673-4680.
**[23]** Turchi, L., Carabelli, M., Ruzza, V., Possenti, M., Sassi, M., Peñalosa, A., Sessa, G., Salvi, S., Forte, V., Morelli, G. and Ruberti, I. 2013. *Arabidopsis* HD-Zip II transcription factors control apical embryo development and meristem function. Develpoment, 140(10): 2118-2129.
**[24]** Vernoud, V., Laigle, G., Rozier, F., Meeley, R. B., Perez, P. And Rogowsky, P. M. 2009. The HD-ZIP IV transcription factor OCL4 is necessary for trichome patterning and anther development in maize. The Plant Journal, 59(6): 883-894.
**[25]** Wilcoxon, F. 1945. Individual comparisons by ranking methods. Biometrics Bulletin, 1(6): 80-83.
**[26]** Zalewski, C. S., Floyd, S. K., Furumizu, C., Sakakibara, K., Stevenson, D. W. and Bowman, J. L. 2013. Evolution of the class IV HD-zip gene family in Streptophytes. Molecular Biology and Evolution, 30(10): 2347-2365.
**[27]** Zhao, Y., Zhou, Y., Jiang, H., Li, X., Gan, D., Peng, X., Zhu, S. and Cheng, B. 2011. Systematic analysis of sequences and expression patterns of drought-responsive members of the HD-Zip gene family in maize. PLoS One, 6(12): e28488.