# Plant Cenh3 Evolution Is Congruent With The Phylogeny Of Plant Species

**Archana Pal, Vishal Singh Negi**

**Abstract**: Centromere plays a major role in the faithful segregation of chromosomes during cell division. This task is achieved by a large protein complex called kinetochore, which is made of several proteins. The centromere is characterized by a histone H3 variant popularly called CENP-A in humans and CenH3 in plants. CenH3 is one of the most rapidly evolving proteins, which is a paradoxical situation for a protein involved in essential biological function. Additionally, many of the kinetochore proteins found in mammals are missing or have extremely high divergence in plants. Therefore, understanding the phylogeny of CenH3 in plants is important for studying kinetochore assembly in plants. In this study, we utilized a computational approach using R and Bioconductor for a comprehensive study of plant CenH3. We found five major clades of plant CenH3 among which the N-terminus is highly divergent and the conserved regions were clustered in three domains. This study has revealed the detailed analyses of plant CenH3 and it will be useful for further investigation aiming at the determination of precise biological functions including its interaction with other proteins that help in the maintenance of centromere structure and function in plants.

**Index Terms**: CenH3, histones, centromere, plants, kinetochore, Bioconductor, Biomart, Ensembl.

## 1. INTRODUCTION

CHROMOSOMES are the functional unit of inheritance, and therefore their proper and faithful segregation to daughter cells is for successful chromosome transmission at cell division [1], [2]. This faithful segregation of chromosomes is ensured by complex biological machinery, the kinetochore, which is formed by the interaction of a large number of proteins on the centromere [3]. The centromere region of a chromosome is characterized by a histone H3 variant known as CenH3 [4]. Usually, the proteins for essential functions in a system are highly conserved. However, despite involvement in the essential biological function, CenH3 proteins are poorly conserved and are known to be rapidly evolving protein [5], [6]. On the other hand, all the core histones including canonical H3 are one of the most conserved eukaryotic proteins [6]. The mammalian centromeric proteins, including CenH3 and its molecular partners, which are involved in the assembly and maintenance of kinetochore in the centromere region are well studied and known [7], [8], [9]. However, their homologs in plants are usually missing. This is conceivable as the CenH3 between mammals and plants is also not conserved and displays a high level of divergence. The study of centromeric proteins in plants has left far behind compared to the same in mammals and as of now, the majority of molecular partners of cenH3 and kinetochore proteins are still not known in plants. This prompted us to undertake a phylogenetic study of CenH3 proteins from plants and to test if their phylogeny is congruent with the phylogeny of plant species.


- *Archana Pal is currently working as an Assistant Professor at PP Savani University, Gujarat, India, PH-+91-6355460976. E-mail: archana.negi@ppsu.ac.in or apal@hawaii.edu*
- *Vishal Singh Negi is a Ford Fellow and is currently working as an Assistant Professor at PP Savani University, Gujarat, India, PH-+91-6355460976. E-mail: vishal.negi@ppsu.ac.in or negi@hawaii.edu*


## 2 METHODS

### 2.1 Software and Programming language used

For this in silico study, the Bioconductor [10], R programming language [11] based open source and open development software project was used for the analysis and comprehension of data. The complete analysis in this study was performed in the RStudio [12], which is the most widely used integrated development environment (IDE) for R. The list of plant datasets was retrieved using the 'listMarts', 'useMart', and 'listDatasets' functions of 'BiomaRt' package [13], [14]. Self-made R scripts in the Bioinformatics laboratory of the PP Savani University, Surat Gujarat were used for all the analysis.

### 2.2 Taxonomic classification of Ensembl plant datasets

The taxonomic classification of each species in the datasets to their respective family and genus were performed using the 'tax_name' function of taxize package [15].

### 2.3 Retrieval of HMGA orthologs from plants

The Oryza sativa japonica was used as the reference plant for the analysis of CenH3 because it is one of the most widely studied plant for which the complete genome is also available. The Ensembl_gene_id for Oryza sativa japonica Os05g0489800 was used to retrieve the complete protein sequence for CenH3 protein. The 'getBM' function of biomaRt was used to extract ensembl_peptide_id and peptide sequence for CenH3 from Oryza sativa dataset (osativa_eg_gene). The ensembl_peptide_id of the corresponding homologs of each available plants listed in the attribute feature of Oryza sativa mart of BiomaRt were obtained from the biomaRt using the getBM function.

### 2.4 Multiple sequence alignment and Phylogenetic analysis of plant CenH3

After retrieving the ensembl_peptide_id for the Oryza sativa japonica CenH3 orthologs, the corresponding sequences were retrieved using biomaRt package (S4, Supplementary materials). The ClustalW multiple sequence alignment was performed in R Studio using the msa package [16]. The distance matrix was computed using 'dist.alignment' function of sequin package [17]. The resultant distance matrices were utilized for constructing a Phylogenetic tree using the neighbor-joining method in the ape package [18], ggplot2

package [19], and ggtree package [20].

## 3 RESULTS

### 3.1 Taxonomic classification of Ensembl plant datasets

The in silico analysis of plant marts in Ensembl biomaRt database identified a total of 59 datasets each for a different plant species (Table S1, Supplementary materials). Analysis of taxonomic classification using taxize package resulted in the identification of 24 different families, namely, 'Actinidiaceae', 'Amborellaceae', 'Apiaceae', 'Asteraceae', 'Bathycoccaceae', 'Brassicaceae', 'Chenopodiaceae', 'Chlamydomonadaceae', 'Cyanidiaceae', 'Cyanidiaceae', 'Dioscoreaceae', 'Fabaceae', 'Funariaceae', 'Gigartinaceae', 'Malvaceae', 'Musaceae', 'Pleosporaceae', 'Poaceae', 'Ranunculaceae', 'Rosaceae', 'Salicaceae', 'Selaginellaceae', 'Solanaceae', and 'Vitaceae'. The taxonomic classification including species, family, genus, and source database for each plant datasets are shown in Table S2 (Supplementary materials).

### 3.2 Retrieval of CenH3 orthologs from Plants

The analysis of plant Ensembl in biomaRt resulted in a total 59 datasets each for a different plant species (Table S1, Supplementary materials). Among all these 59 species, the Oryza sativa japonica was selected as the reference species for two reasons- (i) it is one of the most widely studied species, and (ii) its complete genome has already been sequenced and annotated in greater detail. The retrieved CenH3 protein was found to be 164 aa in size with ensemble_peptide_id Os05t0489800-01 (Table S3, Supplementary materials). As Oryza sativa CenH3 was the reference for which we were retrieving orthologs, the osativa_eg_homolog_ensembl_peptide was removed from the list of attributes for retrieving orthologs from plants. Out of the remaining 58 species listed in datasets, the CenH3 only 44 were identified to have ensembl_peptide_id, rest 14 did not exhibit any ensembl_peptide_id for cenH3. Among the remaining 44, some exhibited only one ensembl_peptide for Oryza sativa CenH3 orthologs while for others multiple IDs were retrieved (Table S4, Supplementary Materials). A total of 52 unique ensembl_peptide_ids were retrieved from 42 plant species (Table S4: Supplementary materials). The protein sequences for each of the 52 unique ensembl_peptide_id for Oryza sativa japonica CenH3 were retrieved using biomaRt. The retrieved protein sequences for 5 ensembl_peptide_id including, AET1Gv20727100.5, AET1Gv20727400.1, HORVU1Hr1G068770.1, HORVU0Hr1G023660.1, OIW14720, appeared partial as the initiator methionine was found missing in these sequences (Table S4, Supplementary Materials). Therefore, these sequences were removed from further analysis and only 47 sequences were used.

### 3.3 Five major evolutionary clades of plant CenH3

Phylogenetic analysis of plant CenH3s (orthologs of Oryza sativa japonica CenH3) resulted in the formation of five major evolutionary clades suggesting their evolution among different families of plants (Fig. 1 and 2).

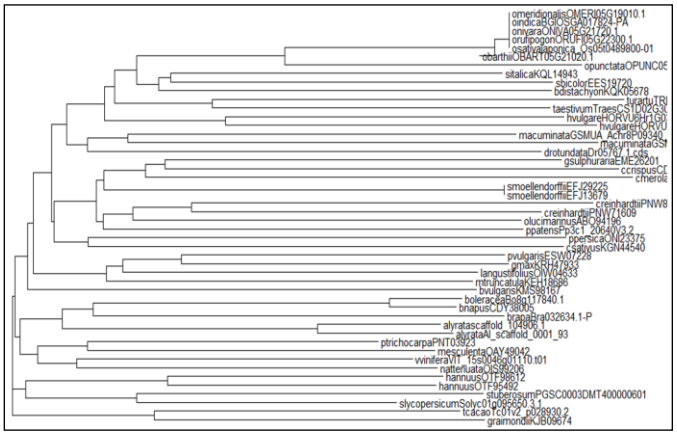

***Figure 1**: Phylogram plot of plant CenH3 proteins.*

Phylogenetic relationship was estimated using neighbor-joining method. Tips of the branches were labeled with the species and ensembl_peptide_id for easy identification of the species and its corresponding peptide id.

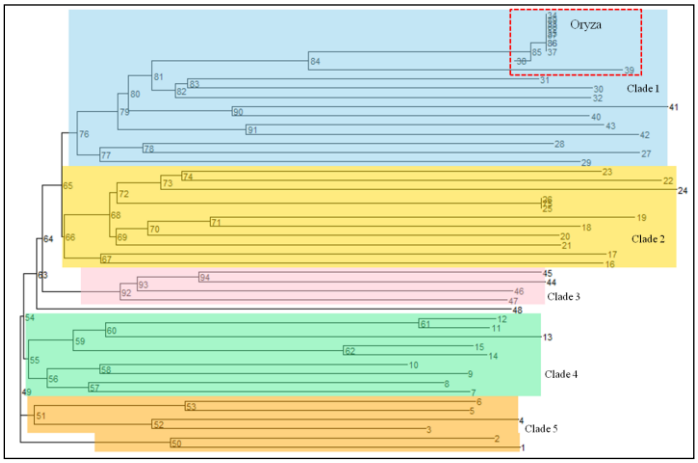


***Figure 2:** Phylogram plot of plant CenH3 and their major clades. Phylogenetic relationship was estimated using neighbor-joining method. Five major evolutionary clades-clade 1, clade 2, clade 3, clade 4 and clade 5 are highlighted in five different colors lightblue, gold, pink, seagreen2, and orange, respectively. The cluster of Oryza branches are shown inside a dotted rectangular box.*

The clade 1 (shaded light blue in Fig. 2) represent species from the family Poaceae such as oryza, sorghum, setaria, brachypodium. The clade 2 comprises of plants including Galdieria sulphuraria, Chondrus crispus, Cyanidioschyzon merolae, Selaginella moellendorffii, Chlamydomonas reinhardtii, Ostreococcus lucimarinus, Physcomitrella patens, Cochliobolus sativus, which belongs to families Cyanidiaceae, Gigartinaceae, Cyanidiaceae, Selaginellaceae, Chlamydomonadaceae, Bathycoccaceae, Funariaceae, Pleosporaceae, respectively. This indicates that clade 2 primarily consists of a mixture of families. The clade 3 appeared as a small group of plants representing Phaseolus vulgaris, Glycine max, Lupinus angustifolius, and Medicago truncatula, all of which, belongs to Fabaceae family. The clade 4 includes plants from Brassicaceae (Brassica oleracea, Brassica napus, Brassica rapa, Arabidopsis lyrata), Salicaceae

(Populus trichocarpa), Vitaceae (Vitis vinifera), and Solanaceae (Nicotiana attenuata) family. The last clade (clade 5) includes plants from Asteraceae (Helianthus annuus), Solanaceae (Solanum lycopersicum, Solanum tuberosum), and Malvaceae (Theobroma cacao, and Gossypium raimondii) families.

### 3.4 Plant CenH3 has three clusters of conserved residues

The ClustalW alignment of plant cenH3 used for constructing the phylogenetic tree was also analyzed for the conserved domains. The alignment was plotted along with the tree and the positions in the alignment were colored based on amino acid type. Interestingly, we identified that the conserved residues were clustered in three different domains (Fig. 3). The levels of conservation were highest in the cluster 3 at the C-terminus end, followed by cluster 2 and then cluster 1. Interestingly, cluster 3 appeared to be conserved more or less uniformly across the clades, whereas cluster 1 and 2 exhibited variation in the degree of conservation which appeared higher to lower from clade 1 to 5, and lower to higher in clades 5 to 1.

## 4. CONCLUSION

The CenH3 protein is the crucial player in the centromere maintenance and function, despite this, it displays high divergence in the primary amino acid sequence across the species. This high level of divergence in the CenH3 sequence explains why many of the mammalian kinetochore proteins are also not conserved and most have different counterparts in the plant kingdom. Considering that it is important for the kinetochore assembly, the lack of conservation in the amino acid sequence of CenH3 can only be explained if CenH3, despite high divergence, retains some of the conserved domains. Our study demonstrates the conserved domains in plant CenH3, which were found to be clustered in three different regions. The N-terminus of all the plant CenH3 displayed high divergence in the amino acid sequence. Our study also identified differential degree of conservation and divergence among different clades and across the clades in the plant kingdom.

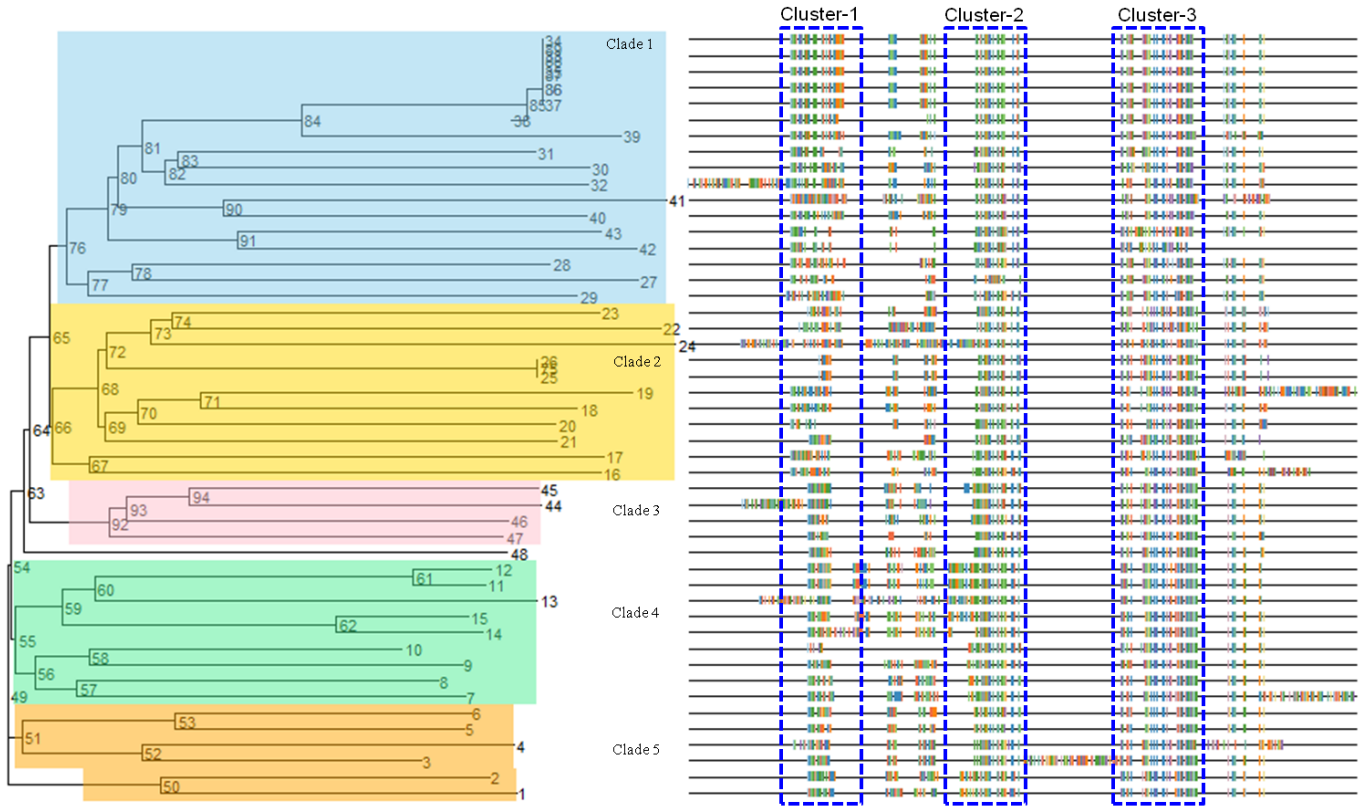


***Figure 3:*** *Phylogram plot and multiple sequence alignment of plant CenH3 proteins.*

Phylogenetic relationship was estimated using neighbor-joining method. Five major evolutionary clades-clade 1, clade 2, clade 3, clade 4 and clade 5 are highlighted in five different colors lightblue, gold, pink, seagreen2, and orange, respectively. The pictorial representation of corresponding ClustalW alignment of CenH3 proteins is shown on the right hand side. The positions in the alignments are colored based on amino acid type. The N-terminus of plant CenH3 displayed high divergence and the conserved regions were found to be clustered in three small regions shown inside the dotted rectangular boxes.

The phylogenetic analysis of plant CenH3 revealed that all the species in clade 1 and clade 3 were from Poaceae and Fabaceae family, suggesting that CenH3 phylogeny in plant is congruent with plant species phylogeny. Therefore, it is conceivable that the other molecular partners of CenH3 or their specific functional domains are also conserved among the plant families. These findings and characterization of plant CenH3 is important for understanding the molecular mechanism of plant kinetochore assembly and also different interaction partners of CenH3, which as of now are poorly studied or known in plants in particular.

## ACKNOWLEDGMENT

VN and AN are thankful to Dr. Parag Sanghani (Provost, PP Savani University) and Shri Vallabhbhai Savani (President, PP Savani University) for their support and providing bioinformatics facility to carry out this study.

## REFERENCES

[1]. Dalal Y, Wang H, Lindsay S, Henikoff S. Tetrameric structure of centromeric nucleosomes in interphase Drosophila cells. PLoS biology. 2007;5(8):e218.

[2]. Kuppu S, Tan EH, Nguyen H, et al. Point mutations in centromeric histone induce post-zygotic incompatibility and uniparental inheritance. PLoS genetics. 2015;11(9):e1005494.

[3]. Jiang J, Birchler JA, Parrott WA, Dawe RK. A molecular view of plant centromeres. Trends in plant science. 2003;8(12):570–575.

[4]. Dalal Y, Furuyama T, Vermaak D, Henikoff S. Structure, dynamics, and evolution of centromeric nucleosomes. Proceedings of the National Academy of Sciences. 2007;104(41):15974–15981.

[5]. Baker RE, Rogers K. Phylogenetic analysis of fungal centromere H3 proteins. Genetics. 2006;174(3):1481–1492.

[6]. Malik HS, Henikoff S. Phylogenomics of the nucleosome. Nature structural & molecular biology. 2003;10(11):882.

[7]. Black BE, Bassett EA. The histone variant CENP-A and centromere specification. Current Opinion in Cell Biology. 2008;20(1):91-100. doi:10.1016/j.ceb.2007.11.007

[8]. Howman EV, Fowler KJ, Newson AJ, et al. Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. PNAS. 2000;97(3):1148-1153. doi:10.1073/pnas.97.3.1148

[9]. Niikura Y, Kitagawa R, Kitagawa K. CENP-A ubiquitylation is inherited through dimerization between cell divisions. Cell reports. 2016;15(1):61–76.

[10]. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology. 2004;5(10):R80. doi:10.1186/gb-2004-5-10-r80

[11]. Ihaka R, Gentleman R. R: A Language for Data Analysis and Graphics. Journal of Computational and Graphical Statistics. 1996;5(3):299-314. doi:10.1080/10618600.1996.10474713

[12]. Racine JS. RStudio: A Platform-Independent IDE for R and Sweave. Journal of Applied Econometrics. 2012;27(1):167-172. doi:10.1002/jae.1278

[13]. Durinck S, Moreau Y, Kasprzyk A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439-3440. doi:10.1093/bioinformatics/bti525

[14]. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature Protocols. 2009;4(8):1184-1191. doi:10.1038/nprot.2009.97

[15]. Chamberlain SA, Szöcs E. taxize: taxonomic search and retrieval in R. F1000Res. 2013;2. doi:10.12688/f1000research.2-191.v2

[16]. Bodenhofer U, Bonatesta E, Horejš-Kainrath C, Hochreiter S. msa: an R package for multiple sequence alignment. Bioinformatics. 2015;31(24):3997-3999. doi:10.1093/bioinformatics/btv494

[17]. Charif D, Lobry JR. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M, eds. Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Biological and Medical Physics, Biomedical Engineering. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007:207-232. doi:10.1007/978-3-540-35306-5_10

[18]. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20(2):289-290. doi:10.1093/bioinformatics/btg412

[19]. Ginestet C. ggplot2: Elegant Graphics for Data Analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2011;174(1):245-246. doi:10.1111/j.1467-985X.2010.00676_9.x

[20]. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017;8(1):28-36. doi:10.1111/2041-210X.12628