Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Uncovering natural allelic and structural variants of *OsCENH3* gene by targeted resequencing and in silico mining in genus *Oryza*

Karminderbir Kaur (kaurkb91@pau.edu)
Punjab Agricultural University
Kumari Neelam
Punjab Agricultural University
Jasdeep Singh
Punjab Agricultural University
Palvi Malik
Punjab Agricultural University
Kuldeep Singh
International Crops Research Institute for the Semi-Arid Tropics

Article

Keywords: OsCENH3, Centromeric Histone 3, Centromeres, Allele mining, Resequencing, Structural variants, CENH3 point mutants

Posted Date: August 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1949764/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on January 16th, 2023. See the published version at https://doi.org/10.1038/s41598-023-28053-w.

Abstract

Plant breeding efforts to boost productivity have focused on developing a haploid development pipeline. *CENH3* gene has emerged as a leading player that can be manipulated to engineer haploid induction system. Currently, allele mining for the *OsCENH3* gene was done by PCR-based resequencing of 33 wild species accessions of genus *Oryza* and *in silico mining* of alleles from pre-existing data. We have identified and characterized CENH3 variants in genus *Oryza*. Our results indicated that the majority CENH3 alleles present in the *Oryza* gene pool carry synonymous substitutions. A few non-synonymous substitutions occur in the N-terminal Tail domain (NTT). SNP A/G at position 69 was found in accessions of AA genome and non-AA genome species. Phylogenetic analysis revealed that non-synonymous substitutions carrying alleles follow pre-determined evolutionary patterns. *O. longistaminata* accessions carry SNPs in four codons along with indels in introns 3 and 6. Fifteen haplotypes were mined from our panel; representative mutant alleles exhibited structural variations upon modeling. Structural analysis indicated that more than one structural variant may be exhibited by different accessions of single species (*Oryza barthii*). NTT allelic mutants, though not directly implicated in HI, may show variable interactions. HI and interactive behavior could be ascertained in future investigations.

Introduction

To overcome limitations in rice productivity, plant breeding efforts have focused on developing a haploid development pipeline. In rice, haploids have been produced primarily through anther culture [1]. Several rice varieties have also been developed and released, particularly in Korea and China. These varieties are of japonica type, which are responsive to anther culture. The potential of this technique in indica rice breeding, however, has not been realized because of the recalcitrant anther culture response of indica varieties, and also because it often produces a high frequency of albino plants that eventually die [1–2]. Thus, anther culture is not a genotype-independent method of haploid induction. To be successful any method of haploid induction should lead to a high frequency of haploid induction and be genotype independent. Wheat × Maize system for wheat haploids is a perfect example of this [3]. Chromosome manipulation and elimination of all alien chromosomes in wide hybrids is a pre-requisite to induce a crop haploid. Rice wide crosses with either maize or pearl millet fail to recapitulate this [4]. Thus, other methods that trigger alien chromosome elimination in crosses of rice need to be explored. Work on intraspecifica nd interspecific hybrids has evidenced the experimental links between the loss of CENH3 (Centromere specific Histone 3 variant) and the occurrence of uniparental chromosome elimination [5–6].

Centromeres control genome inheritance; helping mediate precise movement of chromosomes during cell division by attaching to spindle microtubules via the kinetochore protein complex. The centromere is epigenetically specified by CENH3 [7–8]. It replaces canonical H3 in centromere-specific nucleosomes and recruits various kinetochore proteins [9]. It has been noted that most, but not all centromere histone octamers are populated by CENH3. Conventional histones are a class of highly conserved proteins. On the contrary, CENH3 is rapidly evolving; the tail domain evolves so rapidly that its sequence can barely be aligned between closely related species [10]. The coming together of genomes loaded with centromeres carrying different CENH3 proteins impedes chromosome segregation during zygotic mitosis. The chromosomes carrying ‘defective’ CENH3 are unable to segregate properly and are lost in subsequent cell divisions leading to the production of haploids [5, 10–13]. Thus, the modified *CENH3* gene acts as a haploid inducer gene. It has been reported that progressive natural divergence in CENH3 can cause genome elimination in *Arabidopsis thaliana* [14]. These findings imply that *Arabidopsis* plants having the *cenh3-1* null mutant (embryo lethal) can be complemented with CENH3 from another species or genera to obtain a haploid inducer. Similarly, it has been found that a single point mutation in native CENH3 is sufficient to generate a haploid inducer line [15].

The *CENH3* gene has been functionally characterized in many agronomically important plant species, including rice [16]. In rice, this gene is located on chromosome 5 (http://rice.plantbiology.msu.edu/, Locus ID: LOC_Os05g41080). The natural variation for this has not been exploited in the *Oryza* species. evolutionary and expression patterns of CENH3 in few diploid and allotetraploid species have been evaluated earlier [17]. Protein and nucleotide sequences for CENH3 from various *Oryza* species are available (http://www.uniprot.org/uniprot/; http://www.ebi.ac.uk/ena). Preliminary BLAST analysis of the protein revealed that some of these are identical among species (e.g., *O. latifolia* protein ID E2GIB7_9ORYZ and *O. alta* protein ID D0EKP9_9ORYZ) while others show variation at amino acid level. It is not yet known if this variation may contribute to haploid induction. Uncovering of this knowledge can provide leads towards specific structure of *OsCENH3* and how it can be tinkered to obtain CENH3-based HI lines. In the current study, a detailed analysis of sequence variation at the *OsCENH3* locus (LOC_Os05g41080) was performed *in silico* and in a panel of wild *Oryza* species accessions to identify the evolutionary patterns of *CENH3* alleles. We have identified and characterized *CENH3* variants in genus *Oryza* along with phylogenetic relationships among various *Oryza* species. Our results indicate that the majority *CENH3* alleles present in the *Oryza* gene pool carry synonymous substitutions. A few non-synonymous substitutions where present effect the protein conformation to a certain extent. The phylogenetic analysis also revealed that non-synonymous substitutions carrying alleles do not deviate from expected clustering behavior.

Results

In silico analysis confirms that rice genome has no other paralogs of the OsCENH3 gene

The tBLASTn analysis revealed significant hits on the negative strand of Chromosome 5 in the region 24070193–24067694 which coincided with the *OsCENH3* coordinates (Supplementary Figure S1, Supplementary Table S1) [18]. Similar results were seen upon carrying out Nucleotide BLAST (Supplementary Table S2) [18]. The lack of significant hits on any other part of the genome assembly established that the rice genome does not have any paralogs for the *CENH3* gene.

Upon using the *CENH3* CDS sequences belonging to genus *Oryza* for Selecton analysis, it was seen that the protein was under purifying selection (Fig. 1) [19]. Varying degree of purifying selection was observed among 164 residues of rice CENH3 protein. On the other hand, including the CDS of the *CENH3* gene from various species belonging to genera other than *Oryza*, reflected a slightly different picture (Fig. 2). In the analysis based on a query set of 98 CDS sequences,

many residues of NTT showed positive selection. This implied that the CENH3 protein sequences are highly conserved within the *Oryza* genus, while exhibiting some variation at level of NTT among different genera. This could be explained by the fact that different species and genera have a propensity to incorporate certain residues at higher rates than others. NTT is known to be more gapped/unalignable while HFD is highly conserved [15]. This is also in conference with results obtained earlier by other workers [17].

Allelic variation in the wild species accessions for the gene OsCENH3

Upon amplification with overlapping primer pairs (Supplementary Table S3, Supplementary Figure S3), single sharp bands of expected amplicon size were obtained from the germplasm panel (Supplementary Figure S4, Table 1). The complete allelic sequences of *OsCENH3*, comprised of seven exons (Supplementary Figure S2), could be sequenced in 25 out of 33 wild *Oryza* accessions and were found to be ~ 2166 bp long. In total, *CENH3* alleles from eight *O. rufipogon* accessions, five *O. nivara* accessions, four accessions of *O. glaberrima*, three of *O barthii*, two each of *O. meridionalis* and *O. longistaminata* and one accession of *O. glumaepatula* could be fully sequenced. For the eight accessions where the complete gene sequence could not be obtained, the sequence region pertaining to HFD was analyzed. SNPs observed in exonic regions and major indels found in introns are described in Table 2. No indels were found in exons.

Table 1
List of wild species accessions selected for allele mining of *OsCENH3* locus

Field ID	Species (Accession No.)	Genome	Country of origin
W1	*O. glaberrima (IR100983)*	AA	Senegal
W2	*O. glaberrima (IR101800)*	AA	Nigeria
W3	*O. glaberrima (IR102196)*	AA	Liberia
W27	*O. glaberrima (IR102925)*	AA	Burkina Faso
W39	*O. glaberrima (IR103990)*	AA	Tanzania
W44	*O. barthii (IR89146)*	AA	Zambia
W45	*O. barthii (IR100223)*	AA	Guinea
W46	*O. barthii (IR100934)*	AA	Mali
W53	*O. barthii (IR104102)*	AA	Chad
W58	*O. barthii (IR104136)*	AA	Cameroon
W79	*O. nivara (IR80547)*	AA	India
W82	*O. nivara (IR80722)*	AA	Mynamar
W101	*O. nivara (IR104650A)*	AA	Thailand
W103	*O. nivara (IR104688)*	AA	Sri Lanka
W277	*O. nivara (CR100373)*	AA	India
W509	*O. rufipogon (IR80433)*	AA	India
W517	*O. rufipogon (IR80762)*	AA	Mynamar
W522	*O. rufipogon (IR103404)*	AA	Philippines
W525	*O. rufipogon (IR104389)*	AA	Bangladesh
W526	*O. rufipogon (IR104404)*	AA	Thailand
W634	*O. rufipogon (CR100055)*	AA	India
W726	*O. rufipogon (IR93076)*	AA	Cambodia
W776	*O. rufipogon (IR101411)*	AA	Australia
W924	*O. longistaminata (IR81965)*	AA	Zambia
W926	*O. longistaminata (IR101210)*	AA	Ivory Coast
W927	*O. longistaminata (IR104301)*	AA	Gambia
W944	*O. longistaminata (IR86485)*	AA	Botswana
W950	*O. longistaminata (IR92607)*	AA	Mali
W1014	*O. longistaminata (IR105198)*	AA	Ethiopia
W1031	*O. meridionalis (IR101146)*	AA	Australia
W1037	*O. meridionalis (IR93266)*	AA	Indonesia
W1056	*O. glumaepatula (IR100184)*	AA	Cuba
W1120	*O. australiensis (IR105270)*	EE	Australia

Notably, various SNPs were found across genus *Oryza* irrespective of species E.g., SNP A/G at position 69 was found in accessions of AA genome species (*O. glaberrima, O barthii, O. nivara, O rufipogon, O. longistaminata, O. glumaepatula* and *O. meridionalis*) as well as non-AA genome species (*O. australiensis)*. Some accessions of the progenitor species exhibited no SNP at this position compared to *OsCENH3* which indicated that the actual progenitor of the present-day cultivated rice might have carried this version of the gene. It is interesting to note that synonymous SNPs seen in *O.glaberrima* are also present *O. barthii* (Table 2). Non-synonymous SNPs were found to occur in NTT region only. Two out of the three non-synonymous SNPs were found to exist in different accessions as part of different alleles but were seen to be limited to particular species only (Tables 2 and 3). Based on sequence data available, *O. nivara*, *O.meridionalis*, *O. glumaepatula* and *O. australiensis* (Incomplete sequence in *O. australiensis*) do not carry any non-synonymous SNPs (Table 2).

Table 2
Variations observed in alleles of *CENH3* in wild species of *Oryza*

Variation	Codon/position (ATG)	Effect	Species
GAA > GAG	23/69	Syn	*O. glaberrima, O. barthii, O. nivara, O. rufipogon, O. meridionalis*,
GGC > GGG or GGT	36/233	Syn	*O. glaberrima, O. barthii, O. meridionalis, O. glumaepatula*
ACG > GCG	38/236	A > T	*O barthii*
GGT > GGG	39/241	Syn	*O. glaberrima, O. barthii, O. longistaminata*
AAG > GAG	61/964	K > E	*O. rufipogon*
AGG > ACG	63/971	R > T	*O. longistaminata*
GCA > GCG	74/1005	Syn	*O. nivara, O. australiensis, O. rufipogon*
ACC > ACT	85/ 1038	Syn	*O. glaberrima, O. barthii, O. nivara*,
TCT > TCC	95/1068	Syn	*O. longistaminata*
TGC > TGT	138/1728	Syn	*O. longistaminata*
GCA > GCT	139/1731	Syn	*O. meridionalis, O.glumaepatula*
GCC > GCT	155/ 2136	Syn	*O. longistaminata*
ATC > ATT	158/2145	Syn	*O. longistaminata*
Indel in intron 3	(39 bases)	-	*O. longistaminata*
Indel in intron 6	(38 bases)	-	*O. longistaminata*

Species specific variation in OsCENH3 **alleles in genus** Oryza

It was observed that a few of the SNPs or Indels were species specific (Table 2). Most accessions belonging to *O. longistaminata* manifested a particular haplotype carrying SNPs which were not seen in any accessions belonging to other species of *Oryza* under study. This included SNPs in codons 95, 138, 155 and 158 as well as indels in introns 3 and 6. In a similar trend, SNP at position 1731 (Codon 139) was observed in only *O. meridionalis* and *O.glumaepatula* accessions (Table 2). When compared among different species, the differences arose from major indels in introns (seen in *O. longistaminata* consistently), or different SNPs at the same location. For instance, at position 233, *O. glaberrima* accessions carried base G in place of C in the reference while *O. nivara* as well *O. meridionalis* accessions had a T nucleotide.

Haplotype detection

Fifteen haplotypes (Table 3, Table 4) were detected upon manual curation of SNP data from complete genic sequences or sequences for which sequence data for all exons and relevant introns was available from various accessions. All the 15 haplotypes were comparable to one another at one position or the other. Thus, a single SNP may be part of a number of different haplotypes based on what other SNPs it co-exists with, in different accessions. The accessions belonging to same species tend to have a similar haplotype structure.

Table 3
Haplotypes based on SNPs from complete *CENH3* genic data

SNP/ Haplotype	Codon 23	Codon 36	Codon 38	Codon 39	Codon 61	Codon 63	Codon 74	Codon 85	Codon 95	Codon 138	Codon 139	Codon 155	Codon 158	Indel intron 3	Inde intro 6
OsCENH3	GAA	GGC	ACG	GGT	AAG	AGG	GCA	ACC	TCT	TGC	GCC	GCC	ATC	-	-
H1	GAG	GGG	ACG	GGT	AAG	AGG	GCA	ACT	TCT	TGC	GCC	GCC	ATC	-	-
H2	GAG	GGG	ACG	GGG	AAG	AGG	GCA	ACC	TCT	TGC	GCC	GCC	ATC	-	-
H3	GAG	GGC	ACG	GGT	AAG	AGG	GCA	ACT	TCT	TGC	GCC	GCC	ATC	-	-
H4	GAG	GGG	ACG	GGG	AAG	AGG	GCA	ACT	TCT	TGC	GCC	GCC	ATC	-	-
H5	GAG	GGG	GCG	GGT	AAG	AGG	GCA	ACT	TCT	TGC	GCC	GCC	ATC	-	-
H6	GAG	GGT	ACG	GGT	AAG	AGG	GCA	ACT	TCT	TGC	GCC	GCC	ATC	-	-
H7	GAG	GGT	ACG	GGT	AAG	AGG	GCG	ACC	TCT	TGC	GCC	GCC	ATC	-	-
H8	GAG	GGC	ACG	GGT	GAG	AGG	GCA	ACT	TCT	TGC	GCC	GCC	ATC	-	-
H9	GAG	GGC	ACG	GGT	AAG	AGG	GCG	ACT	TCT	TGC	GCC	GCC	ATC	-	-
H10	GAA	GGC	ACG	GGT	GAG	AGG	GCA	ACT	TCT	TGC	GCC	GCC	ATC	-	-
H11	GAA	GGC	ACG	GGT	AAG	AGG	GCG	ACT	TCT	TGC	GCC	GCC	ATC	-	-
H12	GAG	GGC	ACG	GGT	AAG	ACG	GCA	ACT	TCC	TGT	GCC	GCT	ATT	39	38
H13	GAG	GGC	ACG	GGT	AAG	ACG	GCA	ACT	TCC	TGT	GCC	GCC	ATC	39	38
H14	GAG	GGG	ACG	GGG	AAG	ACG	GCA	ACC	TCC	TGT	GCC	GCT	ATT	39	38
H15	GAG	GGT	ACG	GGT	AAG	AGG	GCA	ACT	TCT	TGC	GCA	GCC	ATC	-	-
H16	GAA	GGC	ACG	GGT	AAG	AGG	GCA	ACT	TCT	TGC	GCC	GCC	ATC	-	-

Table 4
List of wild *Oryza* species accessions carrying different haplotypes

Haplotype	Accessions carrying haplotype
H1	Ogl_IR100983, Ob_IR104102
H2	Ogl_IR101800
H3	Ogl_IR102196, On_IR80722, On_IR104688, Or_IR104389, Or_IR104404, Or_IR93076
H4	Ogl_IR103990
H5	Ob_IR100934
H6	On_IR80547
H7	On_IR104650A
H8	On_CR100373, Or_IR80762, Or_IR103404
H9	Or_IR80433
H10	Or_CR100055
H11	Or_IR101411
H12	Ol_IR81965, Ol_IR86485
H13	Ol_IR104301
H14	Om_IR93266,
H15	Oglp_IR100184

Haplotype structure was seen to vary mostly amongst species but sometimes also within the species. Within the species, these differences were attributable to at most one or two SNPs, that in most cases did not lead to any amino acid changes. Thus, different nucleotide haplotypes could be reduced to a single protein sequence. For instance, haplotypes H8 and H10 translate to identical protein sequence. Similar is the case with H12, H13 and H14. When compared among different species, the differences arose from major indels in introns (seen in *O. longistaminata* consistently), or different SNPs at the same location. Case in point, at position 233, *O. glaberrima* accessions carried base G in place of C in the reference while *O. nivara* as well *O. meridionalis* accessions had a T nucleotide.

Phylogenetic relationships of OsCENH3 gene among diverse wild species germplasm of rice

Upon phylogenetic analysis of complete *CENH3* sequences, the accessions analysed clustered into four major clades (Fig. 3). Upon analysis limited to just the HFD, two major clades were formed (Fig. 4) and the African species *O. barthii* and *O. glaberrima* formed a distinct subclade. *CENH3* from *O. rufipogon* accessions cluster closer to *OsCENH3* reference when analyzed with complete gene sequences as well as with only Histone fold domain (Figs. 3 and 4). The other species were distant depending upon their genome group. *O. meridionalis* and *O. longistaminata* are the most distant among the species under study in both cases. These differences though were not reflected in protein sequences as no major amino acid substitutions were seen, except for in NTT region in *O. rufipogon* and *O. longistaminata*. The *O. rufipogon* accessions (CR100373, IR80762, IR103404 and CR100055) that exhibited amino acid changes (though these were not expected to lead to functional changes in CENH3 behavior) still cluster close to *OsCENH3*.

In silico HFD mining of protein sequences from Oryza genus

The protein length varies from 146–170 residues in different *Oryza* species (Table 5); the insertions and deletions leading to length variation being present in the NTT region. For almost all AA species (except *O. barthii*), CENH3 protein is composed of 164 residues. Ten haplotypes were seen in *Oryza* genus upon aligning and analyzing entire protein length (Table 5). In these ten haplotypes, the major differences arise from changes in NTT, while HFD in many variants is identical. Upon limiting the analysis to just the HFD, five variant types were observed (Table 5). It is noteworthy that *O. meyeriana* carries the CENH3 variant (A0A6G1EGA2) with most changes (ten) in the HFD region when compared to the reference protein Q6T367 (*Oryza sativa ssp japonica*), which can be attributed to the fact that it belongs to the tertiary gene pool. Interestingly, eight out of ten SNPs were not observed in any other species. The most common changes (P70S and L117I) appear to always co-exist and were observed only in the CC genome. Another set of amino acid substitutions (I131L and M152I) also co-exist and were seen to occur in BB, DD as well EE genome. AA genome species do not exhibit any changes in the HFD region, indicating that the CENH3 has remained functionally conserved in the primary gene-pool of the genus *Oryza*. Curiously, the *O. barthii* CENH3 sequence on alignment with other AA-genome CENH3 sequences was revealed to be missing exon 2 completely. Resequencing of *O. barthii CENH3* alleles from our panel revealed one accession to carry a SNP in beginning of exon 2 which we inferred to lead to an amino acid substitution. This anomaly is not seen in any other species including *O glaberrima* which is closely related to *O. barthii*.

Table 5
Haplotype characterization of Oryza genus CENH3 protein sequences based off their alignment with rice CENH3

Species	Accession number	Genome	Protein length (amino acid residues)	e-value	%identity	SNPs observed in HFD	Haplotype	Haplotype on basis of HFD
Oryza alta	D0EKP8	CCDD	164	1.90E-100	90.9	I131L, M152I	Hα	H-I
Oryza alta	D0EKP9	CCDD	166	1.00E-104	93.4	P70S, L117I	Hβ	H-II
Oryza australiensis	D0EKQ2	EE	164	4.40E-91	88.5	I131L, M152I	Hγ	H-I
Oryza barthii	A0A0D3G978	AA	146	3.40E-97	89	none	Hδ	H-III
Oryza glumipatula	A0A0E0A0Z6	AA	164	1.80E-114	100	none	Hε	H-III
Oryza latifolia	E2GIB7	CCDD	166	1.10E-104	93.4	P70S, L117I	Hβ	H-II
Oryza latifolia	E2GIB8	CCDD	164	1.90E-100	90.9	I131L, M152I	Hα	H-I
Oryza meridonalis	A0A0E0DTC0	AA	164	1.80E-114	100	none	Hε	H-III
Oryza meyeriana vra. Granulata	A0A6G1EGA2	GG	146	1.90E-91	82	K80R, T84S, D108E, S110T, L114Y, L118V, I131L, M152I, R162H, P163S	Hζ	H-IV
Oryza minuta	DOEKQ0	BBCC	166	3.84E-103	92.2	K83R	Hη	H-V
Oryza minuta	D0EKQ1	BBCC	166	9.24E-104	92.8	P70S, L117I	Hθ	H-II
Oryza nivara	A0A0E0HG58	AA	164	1.80E-114	100	none	Hε	H-III
Oryza officianalis	E2GIB6	CC	166	9.24E-104	92.8	P70S, L117I	Hθ	H-II
Oryza punctata	D0EKQ3	BBCC	166	2.20E-102	96.6	K83R	Hκ	H-V
Oryza punctata	A0A0E0L436	BBCC	170	6.40E-87	91.3	K83R	Hλ	H-V
Oryza rufipogon	A0A0E0PP98	AA	164	1.80E-114	100	none	Hε	H-III
Oryza rhizomatis	D0EKQ4	CC	166	9.24E-104	92.8	P70S, L117I	Hθ	H-II
Oryza sativa ssp indica	B8AZH6	AA	164	1.80E-114	100	none	Hε	H-III

We found synonymous as well non-synonymous SNPs upon resequencing *CENH3* gene in primary gene pool (AA-genome species) of genus *Oryza*. Additionally, sequencing of the single *O. australienesis* (EE) accession did not yield any SNPs. In contrast, the *in-silico* mining of the protein sequences belonging to species from secondary and tertiary gene pools exhibited several amino acid changes specific to particular genome groups. Most allotetraploid species belonging to the *O. officianalis* complex (secondary gene pool) exhibit at most two changes in HFD. It could also be inferred that various allotetraploid species carry the same CENH3 complement E.g. In *O. alta* (CCDD) *and O. latifolia* (CCDD), both the CENH3 copies are identical (D0EKP8 = E2GIB8, D0EKP9 = E2GIB7). Similarly, *O. rhizomatis* (CC) and *O. officianalis* (CC) CENH3s showed 100% identity.

Mining allelic variants in HDRA and SNPSeek panels

Our analysis of HDRA and SNPseek panels a total of nine OsCENH3 variants, with HDRA returning only one (Supplementary Table S4) [20–22]. The HDRA variant allele (rs# 5655822) was found to have a synonymous effect while of the eight variants mined from SNPseek, only one (rs#175193154) occurred in exon 6 and was found to have non synonymous effect(Cysteine to Serine). Of the remaining seven, six were intronic in nature while one was placed in 3' UTR. rs#175193154 was employed for homology modelling and alignment.

Homology modelling and structural alignment

Secondary structures predicted from variants using PsiPred V4.0 exhibit differences from the structure of the reference OsCENH3 (Fig. 5) [23]. The OsCENH3 secondary structure presents six helices with one helix present in NTT. The haplotype variants H5 and H8 also have six helices but also carry a two-residue strand not seen in the reference. H12 variant has six helices in the regions comparable to the reference but these helices differ in length, a similar case is

presented by rs#175193154 (Supplementary Figure S5). *O. barthii* variant Hδ and *O. officinalis* variant Hθ carry only four helices, all of them comparable to corresponding regions in OsCENH3 HFD and amongst themselves. O. meyeriana CENH3 protein (Hζ) shows five helices.

To better understand structural differences, 3D modelling was carried out. The human nucleosome structure containing the histone variant H3.2–3AV1_A exhibited highest percent identity and was employed as template for homology modelling of the reference (Figure S6). Five models were built for each sequence respectively and the model with minimum DOPE score (Discrete Optimized Protein Energy) was selected (Supplementary Table S5). Protein structures for OsCENH3 and of variant representatives H5, H10, H12 as well as Hδ, Hζ and Hθ belonging to different haplotype groups as well as rs#175193154 were superimposed (Fig. 6, Supplementary Figure S7) and analyzed for structural differences, all models showed > 90% residues in most favored region (Supplementary Figures S8, S9 and S10). The protein structure of OsCENH3 displayed 4 helices (αN Helix, α1 Helix, α2 Helix and α3 Helix) and two loops (loop1 and loop2), loop1and α2 Helix define CATD (Fig. 7), characteristic of CENH3 structure as studied by various workers in different species [15. 24–28].

The comparison of 3D structures showed that H5 and H10 are very similar to OsCENH3 with RMSD (Root mean square deviation) of 2.704 and 8.330 respectively (Supplementary Table S6). No major changes in the overall architecture of proteins were observed. All the helix and loops in H5 were identical to OsCENH3, except the position 71 in H5 (red) was part of helix while it is not in OsCENH3 (green) (Fig. 8). H12 and H10 were identical in this aspect to OsCENH3. Gly 37 to Gly 54 was deleted (disordered region) in Hδ *(O. barthii*) but still all helices and loops are conserved. In Hθ (*O.officinalis*), Thr 38 was absent (purple) and an extra triplet of TAA amino acids occured from position 48–50 still the loops and helix were conserved (Fig. 9). In *O. meyeriana* (Hζ), the helix regions were identical but some variation existed in the loops. Gly 27 to Pro 31 region was absent in OsCENH3. Similar to Hθ, Thr38 was missing in Hζ. An extra loop spanning tripeptide Trp 52 to Ala 54 was also seen in Hζ. Although, the wild relatives exhibit natural variants of *OsCENH3* carrying sequence differences but the mutations (SNPs or Deletions) fail to disrupt the major structural components. Additionally, the rs#175193154 variant mined from SNPseek panel exhibited complete superimposition with reference OsCENH3 (DOPE score: -10105.18555).

Discussion

CENH3 mutations have been profiled extensively in *Arabidopsis thaliana* [15, 29–30]. A corresponding resource does not exist for rice CENH3 mutants. The majority of allelic variation at any given locus is expected to occur in the wild germplasm and not the crop itself, due to the unavoidable loss of variation during domestication [31–33]. In the current study, we have endeavored to scan the *Oryza* germplasm accessions for *CENH3* variants with simultaneous *in silico* analysis. This study underlines that the *OsCENH3* gene is a highly conserved, single copy gene. Contrastingly, barley, wheat and rye have two functional *CENH3* genes: α*CENH3* and β*CENH3* [6, 34–35]. Recent reports in maize suggest that it is easier to manipulate haploid induction in genomes carrying single copy of *CENH3* [27].

Wild germplasm presents synonymous as well as non-synonymous variations

To maximize chances of identifying variant alleles, the allele mining subset was selected using “P” procedure [36]; while also taking into account original geographical distribution of the accessions. Despite having distinctive morphology, life-cycle duration and pollination system, the amino acid sequences of CENH3 in the *Oryza* accessions are highly similar. Synonymous changes outnumber non-synonymous variations at the level of the DNA. The absence of non-synonymous changes across HFD in *CENH3*s of genus *Oryza* is explained by earlier findings wherein it has been shown by the way of multiple sequence alignment that HFD amino acid sequence is similar even among different species belonging to different genera [15]. The *in-silico* analyses of the *OsCENH3* gene carried out as part of the present study also indicated in the same direction in coherence with results obtained in rye [35].

On aligning *Arabidopsis thaliana* and rice CENH3 protein sequences, various motifs were seen to be conserved between these two species (Supplementary Figure S11) which was in coherence with the present investigation as well as previous findings [15]. This could be attributed to the fact that CENH3 like other histones, has certain specific motifs that are indispensable to its function. The MARTKH and PGTVAL motifs, that mark the beginning of NTT and HFD respectively in CENH3 protein and thus define its identity, are seen invariably in all species sequences used for alignment by Kuppu and co-workers [15]. While mining alleles for *OsCENH3* in wild species germplasm of *Oryza* genus, not all accessions could be amplified uniformly and for all regions of the gene. Similar problems were faced by other workers while mining alleles for specific genes from *Oryza* germplasm [37–38]. Transferability of the primer pairs between species belonging to the same genus, is not always 100% as observed in *Passiflora* and *Allium*[39–40].

L117I was one of the most common single-point amino acid change observed on *in silico* analysis of *Oryza* genus CENH3 protein sequences. Its corresponding mutant in *A. thaliana*, L130F/ L130I, has been reported to successfully complement *cenh3* null mutant with haploid inducing abilities (HI, 4.8%) [30]. Arabidopsis CENH3 mutants P82S (corresponding to P70S) and A127V (corresponding to L114Y) also exhibit HI capabilities [15]. Notably, Arabidopsis CENH3 variant with A127T is non-complementing[15, 29].

A total of 157 single-point amino acid substitutions (arising from nucleotide transitions) are possible in *OsCENH3*-HFD (unpublished data), but only a handful of these were seen to exist naturally in *Oryza* relatives and none in closely related- AA genome species. The AA genome species might still be carrying SNPs that are not reflected upon translation as seen on resequencing a subset of accessions for *OsCENH3*. Similar studies in banana (*Musa* spp.) and carrot (*Daucus* spp.) also noted high similarity between different species within the genus [41–42]. Our structural analyses point out that these natural variants do not exhibit major variations in conformation.

Various studies have etched out specific roles played by certain important residues found in NTT region of CENH3 in different species [43–45]. Arginine residues in particular seem to play important role in kinetochore recruitment and subsequent interactions [43–44]. Lysine residue in HFD and Serine residue in NTT have also been implicated in controlling certain functions/behaviours [45–46]. In our analysis, we came across variants in which arginine and lysine

residues had been replaced by other amino acids. These variants can be analyzed further for their interactions during kinetochore recruitment. Another allelic variant discovered from SNPseek panel has a serine residue replacing cysteine, and this might also modulate HI behaviour of lines carrying these.

Phylogenetic implications of resequencing and in silico **analysis**

Phylogenetic relationships among different species of genus *Oryza* have been studied based on entire genomes, chloroplast genome and individual genes by various workers previously. Analysis of various nuclear genes carried out established *O. meridionalis* to be most divergent among AA genome species [47, 48]. This is congruent with our phylogenetic findings based on complete *CENH3* sequence, though when only HFD in analysed, *O.longistaminata* is most distant. This could be explained by presence of two long intronic indels in *CENH3* sequences pertaining to *O.longistaminata*. Interestingly, *O. australiensis* (EE genome) HFD sequence clusters with *O. nivara* and *O. rufipgon*. This is understandable in light of sequenced clones of *Adh1* and *Adh2*, soured from various diploid species of genus *Oryza* being identical respectively [48].

Currently, we sampled eight accessions of *O. rufipogon* and five accessions of *O. nivara* that capture the geopgraphical diversity of the two species. Phylogenetic analyses show that they are scattered in two subclades and overlap with each other as well as *O. sativa* reference, which elucidates their close genetic relationships. On each of phylogenetic trees, *O. rufipogon* and *O. nivara* accessions do not cluster distinctly at the species level. This is identical to findings of Zhu and Ge [48] but in contrast to results obtained by Wambugu and co-workers (2015) who observed even *O. rufipogon* accessions to segregate into “Asian” and “Australian” clusters [49]. Currently, the African species *O. barthii* and *O.glaberrima* form a distinct cluster which was also seen upon analysis of chloroplast sequences [49].

Resequenced variant alleles found in *Oryza* species germplasm exhibit amino acid changes in region of N-terminal tail domain. Histone fold domain does not exhibit any functional SNPs and thus any variants that might be useful towards haploid induction pipeline were not seen in accessions forming our allele mining panel. The single copy *OsCENH3* gene could be easily manipulated for creating mutants that can then be tested for HI trait. The variants found in wild species germplasm along with other OsCENH3 mutants sourced from our mutant garden are under analysis in our lab for their HI capabilities.

Materials And Methods

Preliminary in silico **characterization of** OsCENH3 **gene**

Nucleotide and protein sequences of *CENH3* were retreived from NCBI (https://www.ncbi.nlm.nih.gov/) and UniProt (www.uniprot.org), respectively. Of 894 nucleotide entries found on NCBI, the complete coding sequences and protein sequences were retrieved for 286 entries only. Further, 111 sequences (out of 419 returned upon keyword search) were downloaded from UniProtKB. Redundancy was removed from both nucleotide and protein sequence files using CD-HIT suite with identity cut-off of 0.99, leaving behind 107 nucleotide, 106 CDS derived protein sequences and 89 protein sequences from UniProtKB. The non-redundant nucleotide and protein sequences were subjected to domain analysis using CD search at NCBI server. Any sequences with domains other than pfam00125 (histone-specific domain) were not considered further. The two protein sequence files were combined and redundancy check completed again, returning 144 non-redundant protein sequences. Thus, 144 protein and 107 nucleotide sequences were used for investigating any *OsCENH3* paralogs. The sequences were BLAST searched [18] (BLAST + version was used; [50]) against the complete rice genome assembly version 7.0 [51], retrieved from the RGAP database (ftp://ftp.plantbiology.msu.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_7.0/all.dir/). Standalone BLASTn and tBLASTn searches were performed using the rice genome assembly as database and the *CENH3* CDS and protein FASTA files as the query sequences. In a separate analysis, effect of amino acid substitution in the CENH3 sequences was anlysed using Selecton server.

Natural allelic diversity for CENH3 **gene in wild species of rice**

Plant Materials

A set of 33 wild species accessions representing eight diploid wild species of genus *Oryza* viz. *O. rufipogon* (n = 8), *O. nivara* (n = 5), *O. glaberrima* (n = 5), *O. barthii* (n = 5), *O. glumaepatula* (n = 1), *O. longistaminata* (n = 6), *O. meridionalis* (n = 2), *O. australienesis* (n = 1) were undertaken for the present study. These accessions were initially procured from International Rice Research Institute (IRRI), Philippines and National Rice Research Institute (NRRI), Cuttack, India and are being actively maintained at Punjab Agricultural University, Ludhiana. Standard agronomic practices were followed to raise the crop. One-month-old seedlings were transplanted in the field with a row-to-row distance of 60 cm and plant-to-plant distance of 40 cm; weed control was done using a Paddy Weeder, 15 days after transplanting and again after a fortnight. Application of organic manures (15 tonnes of farmyard manure per hectare), along with chemical fertilizers comprising 222 kg/ha Neem coated urea (provide 104 kg/ha Nitrogen), 67 kg/ha Diammonium Phosphate (provide 30 kg/ha phosphorus), and 49 kg/ha Muriate of potash (provide 30 kg/ha potassium) prior to transplanting of rice was carried out for higher crop yield and maintenance of soil health. 1/3 nitrogen was applied within 2 weeks of transplanting while the whole phosphorus and potassium was applied before the last puddling. Broadcasting of the remaining nitrogen was done in two equal splits, one three weeks after transplanting and the other three weeks afterwards; water was kept standing in the crop continuously for two weeks after transplanting, and thereafter irrigation was done two days after the ponded water had infiltrated into the soil; to facilitate harvesting, irrigation was stopped about a fortnight before maturity. Panicles of the plants were covered with the mesh bags to avoid shattering of the seeds; harvesting and threshing of different genotypes was done separately to avoid seed admixture.

DNA isolation, Primer designing and PCR amplification: Genomic DNA was isolated from leaves of selected accessions using a modified CTAB method. 0.8% agarose gel was used to access the quantity and quality of DNA from each sample. To amplify coding and non-coding regions of *OsCENH3* alleles, the complete sequence of chromosome 5 was retrieved and aligned with gene sequence obtained from RGAP database (RGAP Locus ID: LOC_Os05g41080, http://rice.plantbiology.msu.edu//index.shtml) using offline BLAST and further trimmed to obtain gene sequence flanked by additional 267 bp before the start codon and 467 bp after stop codon, yielding a sequence of 2900 bases. This sequence was used to design three sets of overlapping primer pairs

(Supplementary Table S3). Supplementary Figure S1, S2 show structure of the gene and Supplementary Figure S3 shows position of primers along the length of the gene. PCR was performed in a 30 µl reaction mix containing 0.25 µl TaKaRa Ex-Taq DNA polymerase, 1 µl of genomic DNA (100ng/µl), 3 µl of 10X TaKaRa Ex-Taq buffer, 3 µl of dNTPs (1 mM), 1.5 µl each of forward and reverse primers (5 µM), and 19.75 µl Nuclease Free Water. The thermal cycling conditions were as follows: an initial denaturation at 94°C for 5 min; 35 cycles of 1 min denaturation at 94°C followed by 45 sec annealing at 55°C and 1 min extension at 72°C; and a final 7 min extension at 72°C. Detailed protocols may be found in [52].

Sequencing of *OsCENH3* gene in selected accessions

Ethidium bromide stained 1.0% agarose gel electrophoresis was performed to analyze PCR products. 5 µl PCR product for each sample was loaded onto the gel along with 1 kb plus ladder (Thermo Scientific Generuler) to estimate the DNA fragment size. We obtained single sharp bands of expected amplicon size for all the three overlapping primers (Supplementary Figure S4). The Wizard® SV PCR Clean-Up System (Promega, USA) as per the manufacturer's protocol was followed to purify the DNA fragments from remaining 25 µl PCR product. The information on the nucleotide sequence of the target PCR products was generated using the forward and reverse primers, either in-house with the Big dye Terminator v3.1 (ABI Sequencer 3730xl) or by outsourcing purified samples to Agrigenome Labs Pvt. Ltd., New Delhi. Experiment was carried out in two replications to confirm the presence of single nucleotide polymorphism (SNPs).

Analysis of the generated nucleotide sequences: For comparative sequence analysis, contigs were assembled from individual reads produced by overlapping primers, using DNA Baser Assembler v5.15.0., to generate the contiguous sequence of *OsCENH3* alleles from selected genotypes. This software also helps in automatic identification and clipping of poor-quality regions at both ends of the sequences. ClustalX 2.0.11 was employed to align contig sequences individually with the reference (RGAP Locus ID: LOC_Os05g41080) and to trim the sequences at both ends to retain only the genic portion (i.e., from ATG to TGA). Sequences were also trimmed to retain just HFD or individual exons.

Pairwise alignments and multiple sequence alignment of the trimmed sequences with annotated version of the reference were carried out using Geneious Prime, for ease of visualization. Based on the alignment, SNPs and haplotypes were predicted. The detected SNPs were then manually curated by analyzing and comparing chromatogram files to the Geneious alignment files and DNA Baser contig files. Effects of detected SNPs in terms of coded amino acids, were also visualized using Geneious Prime.

Phylogenetic analysis

The MEGA X software was used to generate the phylogenetic tree using multiple sequence alignment file [53]. The evolutionary distances were computed using the Neighbor Joining Method with 1,000 bootstraps using the Kimura2-parameter model.

In silico **mining of** OsCENH3 **alleles**: The *OsCENH3* alleles from wild *Oryza* species not represented in the resequencing panel, were mined *in silico* on UniProt server. Eleven protein sequences (D0EKP8, D0EKP9, D0EKQ2, A0A0D3G978, A0A0EA0Z6, E2GIB7, E2GIB8, A0A0EDTC0, A0A6EGA2, DOEKQ0, DOEKQ1, A0A0E0HG58, E2GIB6, D0EKQ3, A0A0E0L436, A0A0E0PP98, D0EKQ4, B8AZH6 and Q6T367) were retrieved using search "CENH3 AND *Oryza*". Q6T367 is the CENH3 protein from *O. sativa ssp. japonica* pertaining to gene ID Os05g0489800 which is a RAP-DB (https://rapdb.dna.affrc.go.jp/) equivalent of RGAP Locus ID: LOC_Os05g41080 and was thus used as the reference for pairwise alignment of the other 19 sequences using blastp at UniProt server. The alignments were studied to mine SNPs in Histone Fold Domain.

Additonally, SNP calling was carried out on SNPSeek and HDRA panels in the region limited to OsCENH3 coordinates. Effects of detected SNPs in terms of coded amino acids, were also visualized using Geneious Prime.

Structure analysis of variants: Representative variant protein sequences along with OsCENH3 reference were subjected to secondary structure prediction using PsiPred [54–55]. For 3D modelling, OsCENH3 protein sequence was subjected to blastp search against the PDB database at NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) for identification of a suitable template. Homology modeling of 3D structures for OsCENH3 as well as variant representatives H5, H10, H12 as well as Hδ, Hζ and Hθ, and rs#175193154 was done using Modeller 10.1 [56] Quality of structures was accessed through the SAVES server V 6.0 (https://saves.mbi.ucla.edu/) and Ramachandran plot was made for the selected models to assess the stereo-chemical properties. The structures were aligned with OsCENH3 using PyMOL. OsCENH3 was also aligned with 3av1.

Declarations

Acknowledgements

The authors acknowledge that this study was carried out as part of doctoral thesis of Karminderbir Kaur. The authors are grateful to Bayer's Beachell Borlaug International Scholars Program (earlier Monsanto's Beachell Borlaug International Scholars Program) and Council of Scientific and Industrial Research and University Grants Commission administered joint Junior Research Fellowship for financial support. The authors are grateful to Dr. Anne Britt, Dr. Luca Comai, Dr. Nitzan Shabek and Dr. Sundaram Kuppu of University of California, Davis for their inputs and discussions that helped shape this manuscript. The authors also express their thanks to Dr. S.S. Banga, Dr. Parveen Chhuneja, Dr. Yogesh Vikal and Dr. Priti Sharma for their feedback.

Author Contributions

KS, KN and KK conceived the concept and designed the experiment. Material preparation, data collection and analysis were performed by KK, JS and PM. The manuscript was written and edited by KK and KN. All authors read and approved the final manuscript.

Availability of Data and Materials

The sequence data generated as part of this study have been submitted to GenBank, Accession numbers: OK500353-OK500376 are available to public while accession numbers OM514977-OM514984 should be available to public in September 2022. Materials can be requested to KN.

Link for Data: https://www.ncbi.nlm.nih.gov/nuccore/OK500353.1/

Similarly replace GenBank accession numbers (OK500353-OK500376) can be replaced in the above link.

Plant material permission

The materials were initially procured from International Rice Research Institute (IRRI), Philippines and National Rice Research Institute (NRRI), Cuttack, India and are being actively maintained at Punjab Agricultural University, Ludhiana. For official accession numbers please refer to Table 1.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Funding

The present study was funded by financial aid received through Bayer's Beachell Borlaug International Scholars Program (earlier Monsanto's Beachell Borlaug International Scholars Program) and through CSIR-UGC joint Junior Research Fellowship.

References

1. Gueye, T., & Ndir, K. In vitro production of double haploid plants from two rice species (*Oryza sativa L.* and *Oryza glaberrima (Steudt.)* for the rapid development of new breeding material. Sci. Res. Essays **5**, 709–713 (2010)
2. Mishra, R., Rao, G.J.N., Rao, R.N. & Kaushal, P. Development and characterization of elite doubled haploid lines from two *indica* rice hybrids. Rice Sci. **2**, 290–299 (2015).
3. Srivastava, P. & Bains, N.S. Accelerated wheat breeding: doubled haploids and rapid generation advance. In: *Biotechnologies of Crop Improvement* Vol. 1 Pp. 437–461. (Springer, Cham 2018).
4. Kaur, K., Gupta, M., Vikal, Y., Singh, K., & Neelam, K. Callose depositions underlie the incompatible reaction in intergeneric crosses of rice. Plant Genet. Resour. **19**, 447–452 (2021).
5. Ravi, M. & Chan, S.W. Haploid plants produced by centromere-mediated genome elimination. Nature **464**, 615–618 (2010).
6. Sanei, M., Pickering, R., Kumkea, K., Nasudac, S. & Houben, A. Loss of centromeric histone H3 (CENH3) from centromeres precedes uniparental chromosome elimination in interspecific barley hybrids. *Proc. Natl. Acad. Sci.* **108**, e498 - e505 (2011).
7. Earnshaw, W.C. & Rothfield, N. Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. Chromosoma **91**, 313–321 (1985).
8. Talbert, P.B., Masuelli, R., Tyagi, A.P., Comai, L. & Henikoff, S. Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. Plant Cell **14**, 1053–1066 (2002).
9. Henikoff, S. and Dalal, Y. Centromeric chromatin: what makes it unique? Curr. Opin. Genet. Dev. **15**, 177–184 (2005).
10. Ravi, M., *et al*. The rapidly evolving centromere-specific histone has stringent functional requirements in *Arabidopsis thaliana*. Genetics **186**, 461–471 (2010).
11. Copenhaver, G.P. & Preuss, D. Haploidy with histones. Nat. Biotechnol. **28**:423–424 (2010).
12. Chan, S.W. In a battle between parental chromosomes, a failure to reload. *Proc. Nat. Acad. Sci.* **108**, 13361–13362 (2011).
13. Comai, L. Genome elimination: translating basic research into a future tool for plant breeding. PLoS Biol. **12**, e1001876 (2014).
14. Maheshwari, S., Tan, E.H., West, A., Franklin, F.C., Comai, L. & Chan, S.W. Naturally occurring differences in CENH3 affect chromosome segregation in zygotic mitosis of hybrids. PLoS Genet **11**, e1004970 (2015).
15. Kuppu, S., *et al.* Point mutations in centromeric histone induce post-zygotic incompatibility and uniparental inheritance. PLoS Genet. **11**, e1005494 (2015).
16. Nagaki, K., *et al.* Sequencing of a rice centromere uncovers active genes. Nat. Genet. **36**, 138–145 (2004).
17. Hirsch, C.D., Wu, Y., Yan, H. and Jiang, J. Lineage-specific adaptive evolution of the centromeric protein CENH3 in diploid and allotetraploid *Oryza* species. Mol. Biol. Evol. **26**: 2877–2885 (2009).
18. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. Basic local alignment search tool. J. Mol. Biol. **215**, 403–410 (1990).
19. Doron-Faigenboim, A., Stern, A., Mayrose, I., Bacharach, E., & Pupko, T. Selecton: a server for detecting evolutionary forces at a single amino-acid site. Bioinformatics **21**, 2101–2103 (2005).
20. McCouch, S.R., *et al.* Open access resources for genome-wide association mapping in rice. Nature Comm. **7**, 1–14 (2016).
21. Mansueto, L., *et al*. SNP-Seek II: A resource for allele mining and analysis of big genomic data in Oryza sativa. Curr Plant Biol. **7**,16–25 (2016).
22. Mansueto, L., *et al.* Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res. **45**, D1075-D1081 (2017).
23. McGuffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. Bioinformatics **16**, 404–405.

24. Black, B.E., *et al.* Centromere identity maintained by nucleosomes assembled with histone H3 containing the CENP-A targeting domain. Mol. Cell **25**, 309–322 (2007).
25. Black, B.E., *et al.* Structural determinants for generating centromeric chromatin. Nature **430**, 578–582 (2004).
26. Feng, C., *et al.* The deposition of CENH3 in maize is stringently regulated. Plant J. **102**, 6–17 (2020).
27. Wang, N., Gent. J.I. & Dawe, R.K. Haploid induction by a maize cenh3 null mutant. Sci. Adv. **7**, eabe2299 (2021).
28. Zhou, Z., *et al.* Structural basis for recognition of centromere histone variant CenH3 by the chaperone Scm3. Nature **472**, 234–237 (2011).
29. Kuppu, S., et al. A variety of changes, including CRISPR/Cas9-mediated deletions, in CENH3 lead to haploid induction on outcrossing. Plant Biotechnol. J. **18**, 2068–2080 (2020).
30. Karimi-Ashtiyani, R., *et al.* Point mutation impairs centromeric CENH3 loading and induces haploid plants. *Proc. Nat. Acad. Sci.* **112**, 11211–11216 (2015).
31. Tenaillon, M.I., U'Ren, J., Tenaillon, O. & Gaut, B.S. Selection versus demography: a multilocus investigation of the domestication process in maize, Mol. Biol. Evol. **21**, 1214–1225 (2004).
32. Hyten, D.L., Song, Q., Zhu, Y., *et al.* Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Nat. Acad. Sci.* **103**, 16666–16671 (2006).
33. a] Zhu, Q.H., Zheng, X.M., Luo, J.C., Gaut, B.S. & Ge, S. Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: severe bottleneck during domestication of rice. Mol. Biol. Evol. **24**, 875–888 (2007).
34. Yuan, J., Guo, X., Hu, J., Lv, Z., & Han, F. Characterization of two CENH 3 genes and their roles in wheat evolution. New Phytologist, **206**, 839–851 (2015).
35. Evtushenko, E.V., *et al.* Conserved molecular structure of the centromeric histone CENH3 in Secale and its phylogenetic relationships. Sci. Rep. **7**, 1–10 (2017).
36. Reeves, P.A., Panella, L.W. & Richards, C.M. Retention of agronomically important variation in germplasm core collections: implications for allele mining. Theor. Appl. Genet. **124**, 1155–1171 (2012).
37. Kaur, A., *et al.* Novel cis-acting regulatory elements in wild Oryza species impart improved rice bran quality by lowering the expression of phospholipase D alpha1 enzyme (OsPLDα1). Mol. Biol. Rep. **47**, 401–422 (2020).
38. Latha, R., Rubia, L., Bennett, J. & Swaminathan, M.S. Allele mining for stress tolerance genes in Oryza species and related germplasm. Mol. Biotechnol. **27**, 101–108 (2004).
39. Oliveira, G.A.F., *et al.* Cross-species amplification of microsatellite loci developed for Passiflora edulis Sims. in related Passiflora Species. Braz. Arch. Biol. Technol. **56**, 785–792 (2013).
40. Wilkie, S.E., Isaac, P.G. & Slater, R.J. Random amplified polymorphic DNA (RAPD) markers for genetic analysis in Allium. Theor. Appl. Genet. **86**, 497–504 (1993).
41. Muiruri, K.S., *et al.* Expressed centromere specific Histone 3 (CENH3) variants in cultivated triploid and wild diploid bananas (Musa spp.). Front Plant Sci 8: 1034 (2017).
42. Dunemann F, Schrader O, Budahn H, Houben A (2014) Characterization of centromeric histone H3 (CENH3) variants in cultivated and wild carrots (Daucus sp.). Plos One, 9: e98504.
43. Samel A, Cuomo A, Bonaldi T and Ehrenhofer-Murray AE (2012) Methylation of CenH3 arginine 37 regulates kinetochore integrity and chromosome segregation. Proceedings of the National Academy of Sciences. 109:9029–9034.
44. Torras-Llort M, Medina-Giro S, Moreno-Moreno O and Azorin F (2010) A Conserved Arginine-Rich Motif within the Hypervariable N-Domain of Drosophila Centromeric Histone H3 (CenH3CID) Mediates BubR1 Recruitment. PloS One. **5**, e13747.
45. Demidov, D., *et al.* Deregulated phosphorylation of CENH3 at Ser65 affects the development of floral meristems in Arabidopsis thaliana. Front. Plant Sci. **10**, 928–939 (2019).
46. Niikura, Y., *et al.* CENP-A K124 ubiquitylation is required for CENP-A deposition at the centromere. Developmental Cell. **32**, 589–603 (2015).
47. Ge, S., Sang, T., Lu, B.R. & Hong, D.Y. Phylogeny of rice genomes with emphasis on origins of allotetraploid species. *Proc. Nat. Acad. Sci.* **96**, 14400–14405 (1999).
48. Zhu, Q. & Ge, S. Phylogenetic relationships among A-genome species of the genus Oryza revealed by intron sequences of four nuclear genes. New Phytologist. **167**, 249–265 (2005).
49. Wambugu, P.W., Brozynska, M., Furtado, A., Waters, D.L. & Henry, R.J. Relationships of wild and domesticated Rices (Oryza AA genome species) based upon whole chloroplast genome sequences. Sci. Rep. **5**, 13957 (2015).
50. Camacho, C., et al. (2009). BLAST+: architecture and applications. BMC Bioinform. **10**, 1–9.
51. Kawahara, Y., *et al.* Improvement of the *Oryza sativa Nipponbare* reference genome using next generation sequence and optical map data. Rice **6**,4 (2013).
52. Kaur, K. 'Development of an *in vivo* haploid induction system in rice through distant hybridization and manipulation of *CenH3* gene', PhD thesis, Punjab Agricultural University, Ludhiana (2020).
53. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. **35**, 1547 (2018).
54. Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. **292**, 195–202 (1999).
55. Buchan, D.W. & Jones, D.T. The PSIPRED protein analysis workbench: 20 years on. Nucleic Acids Res. **47**, W402-W407 (2019).
56. Webb, B., & Sali, A. Comparative protein structure modeling using MODELLER. Curr. Protoc. Bioinform. 54, 5–6 (2016).

Figures



Figure 1

Selecton analysis of Oryza genus CENH3 gene The Selecton analysis revealed rice CENH3 protein to be under purifying selection.

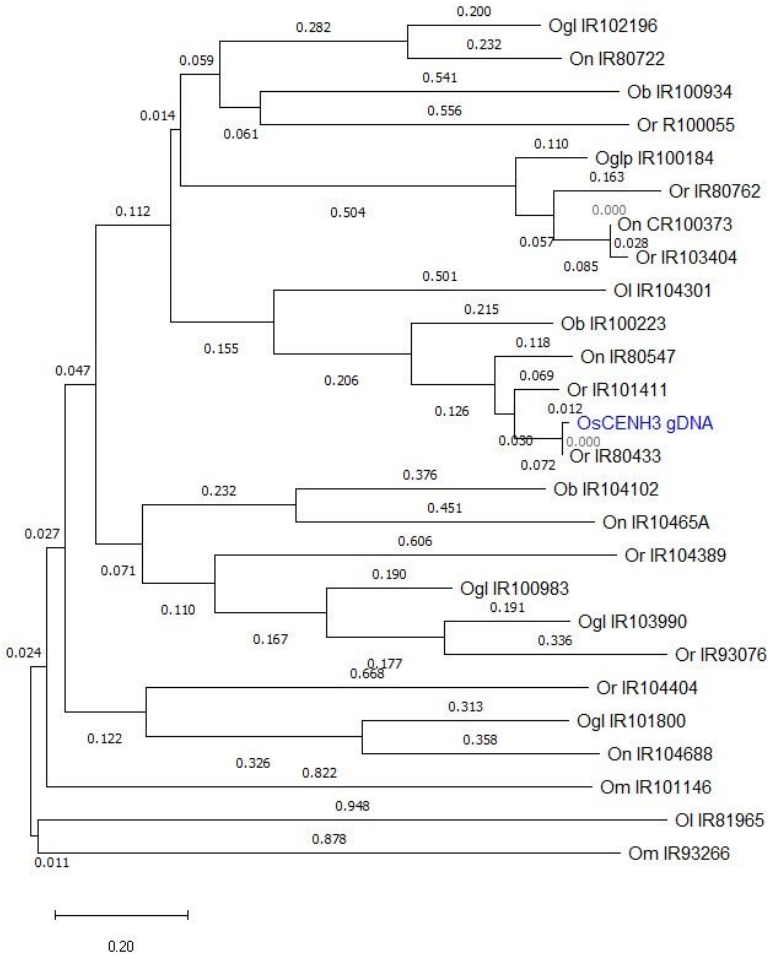


Figure 3

Phylogenetic analysis of resequenced CENH3 alleles Dendrogram generated using complete CENH3 sequences of 24 Oryza accessions in MEGA X at bootstrap 1000 using neighbor-joining method. Individual accessions are named as abbreviated species name followed by international accession number as listed in Table 1. Species abbreviated as: Ob = *O. barthii*, Ogl = *O. glaberrima*, Oglp = *O. glumaepatula*, Om = *O. meridionalis*, On = *O. nivara*, O r= *O. rufipogon*.

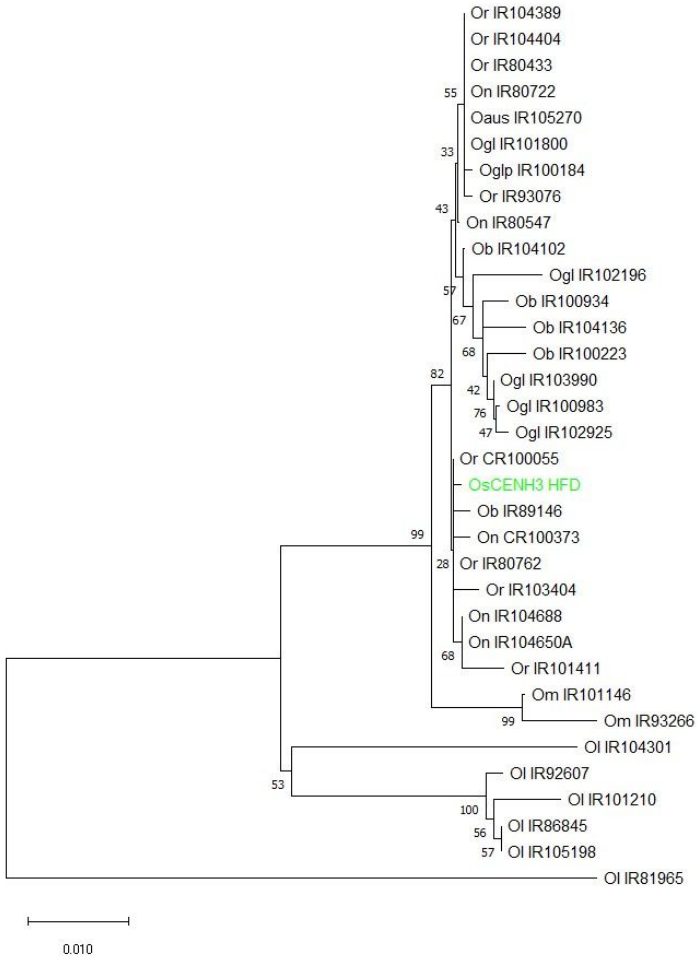


Figure 4

Phylogenetic analysis Limited to HFD domain of resequenced alleles Dendrogram generated using only CENH3 HFD region sequences of 33 accessions in MEGA X at bootstrap 1000 using the neighbour-joining method. Species representation on same basis as Figure 3, additionally Oaus = *O. australiensis*.

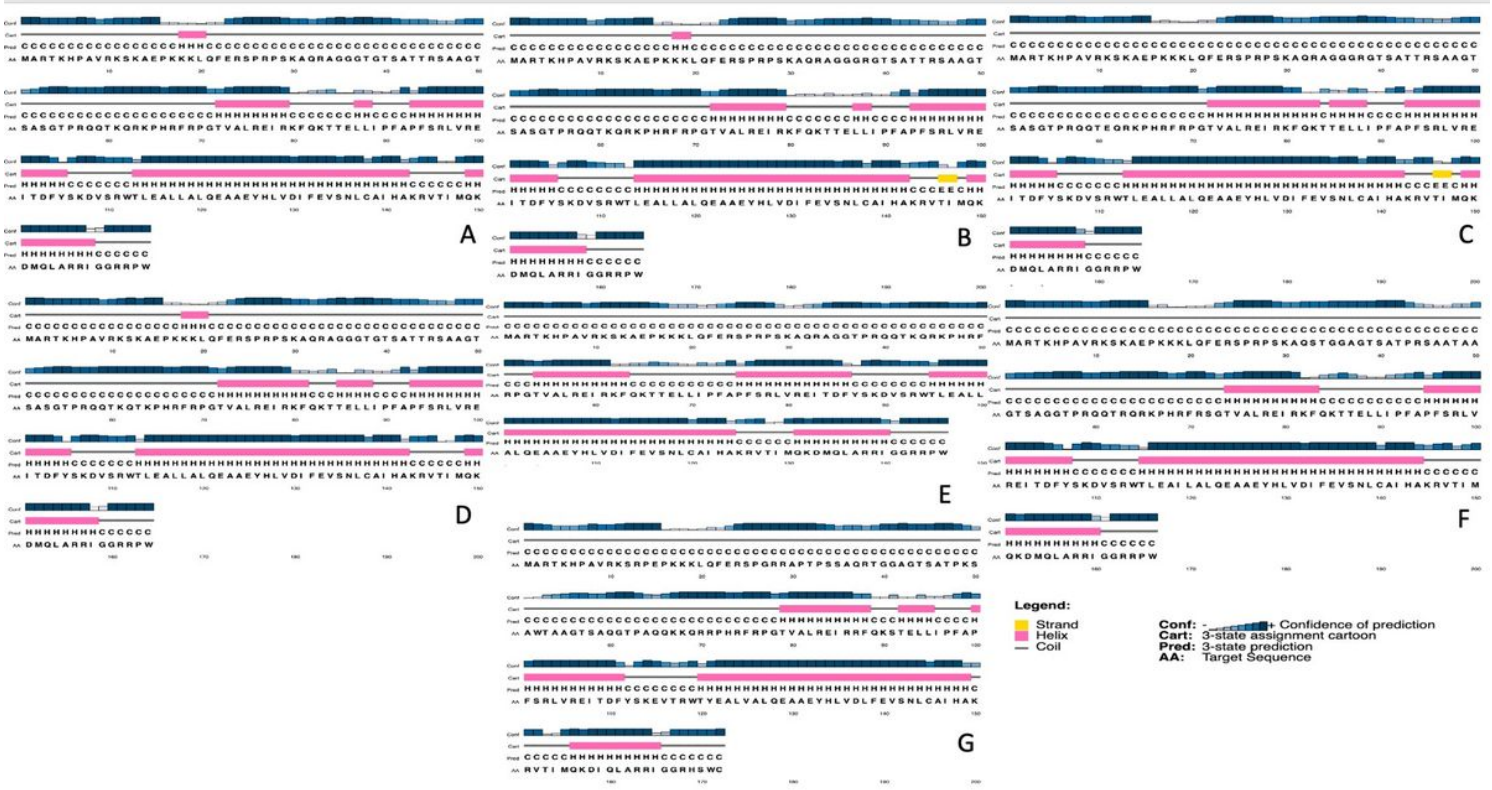


Figure 5

Secondary structure prediction Secondary structure was predicted for A) OsCENH3 reference, B) H5, C) H10, D) H12, E) Hδ, F) Hζ, and G) Hθ and various differences were observed among different alleles as discussed in text.

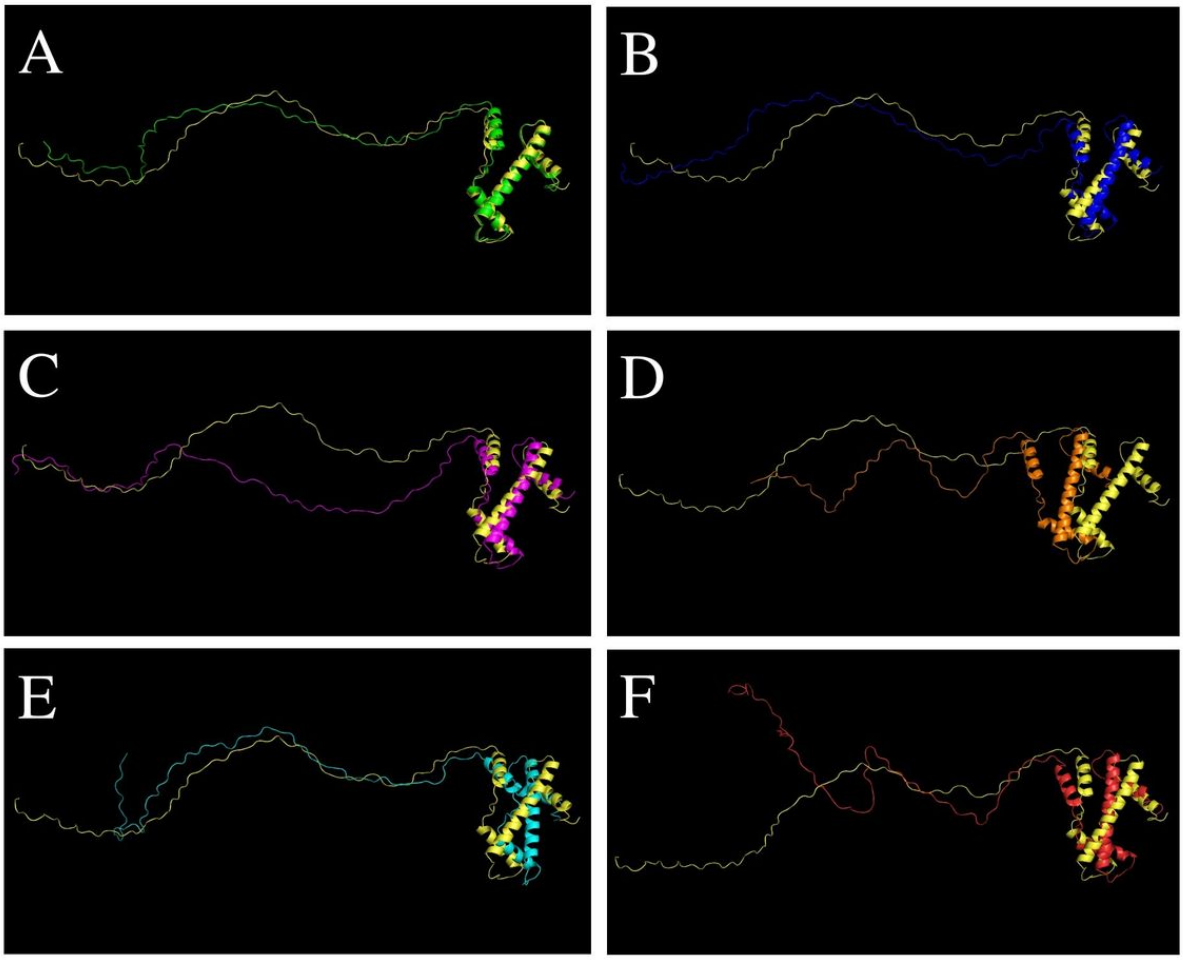


Figure 6

Comparison of 3D models of *Oryza* genus CENH3s Superimposition of reference OsCENH3 (yellow) with A) H5, B) H12, C) H10, D) Hδ, E) Hθ, and F) Hζ.

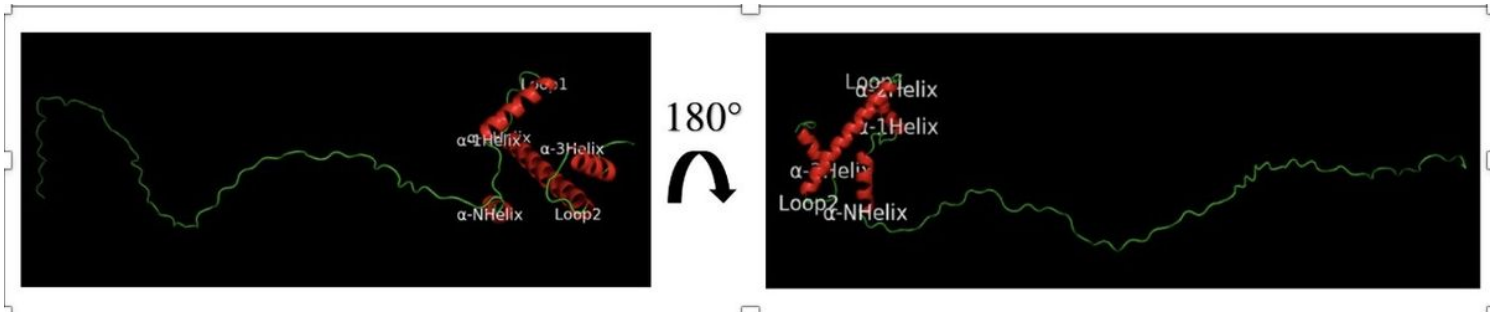


Figure 7

Detailed structure of OsCENH3 OsCENH3 displays 4 helices (αN Helix, α1 Helix, α2 Helix and α3 Helix) and two loops (loop1 and loop2) colored red and green respectively.

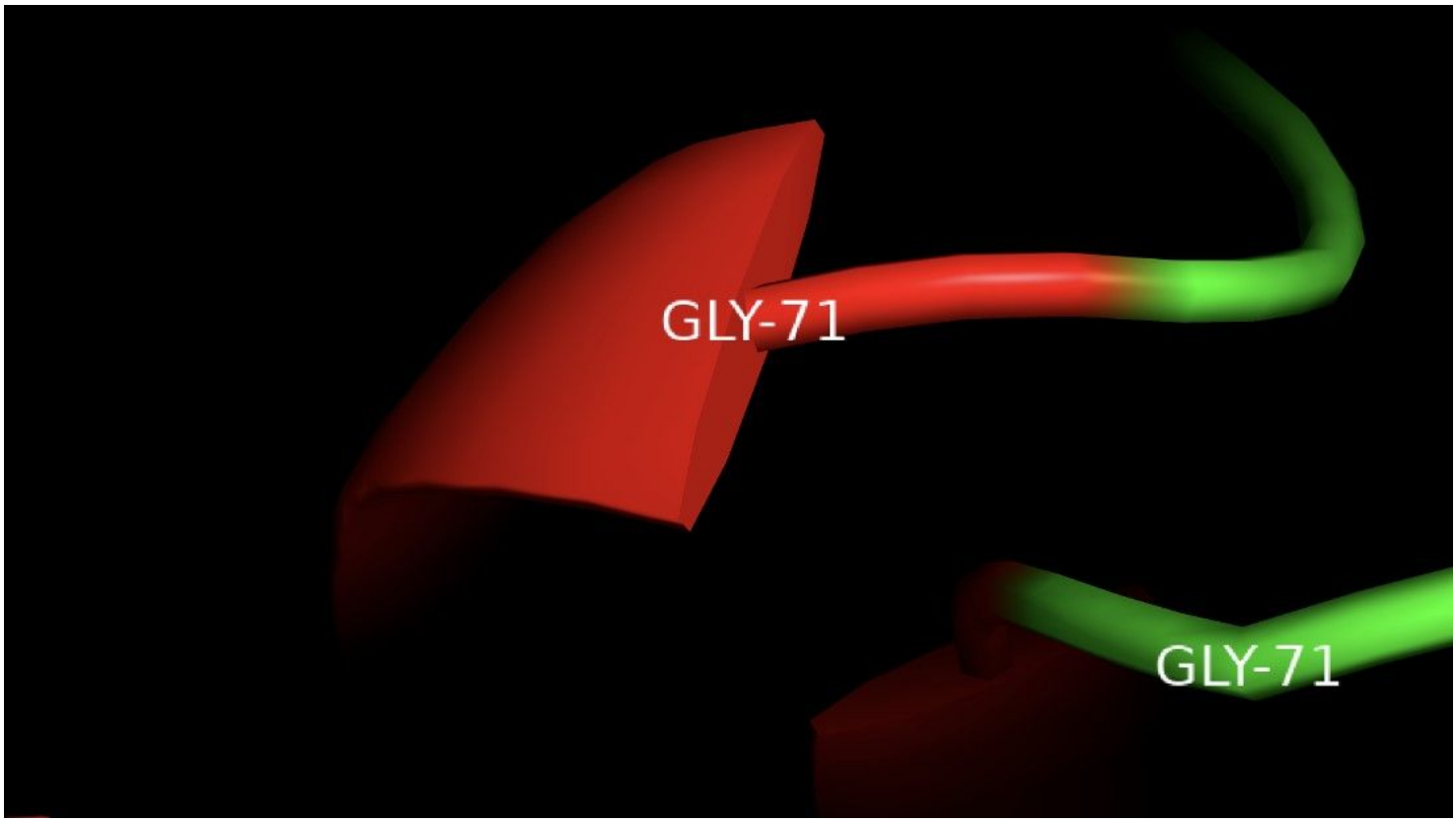


Figure 8

V**ariations in OsCENH3 and H5** Position 71 in H5 (red) is part of helix while it is not in OsCENH3 (green).

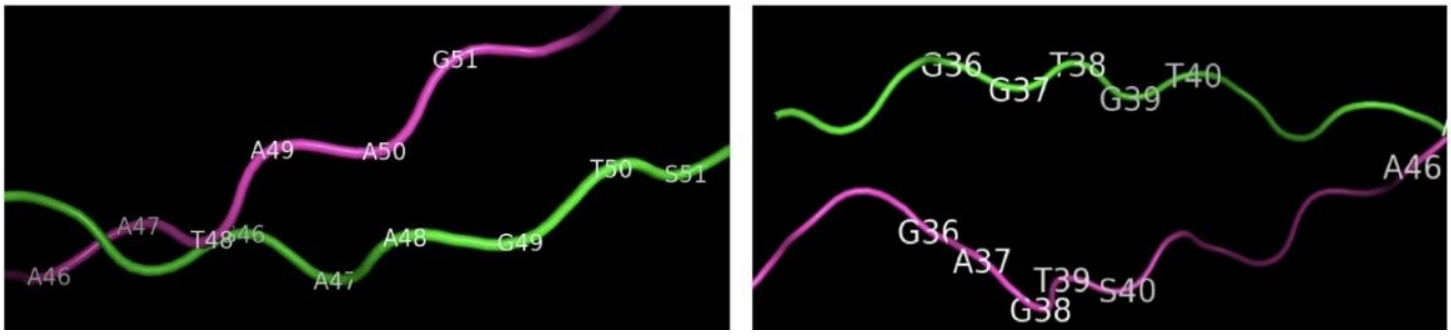


Figure 9

Variations in OsCENH3 and Hθ Thr 38 is absent (purple) and an extra triplet of TAA amino acids occur from position 48-50 in Hθ.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Suppfigures.docx.pdf
- SupplementarytableS1.pdf
- SupplementarytableS2.pdf
- Supplementarytables35and6.docx
- SupplementarytableS4.pdf