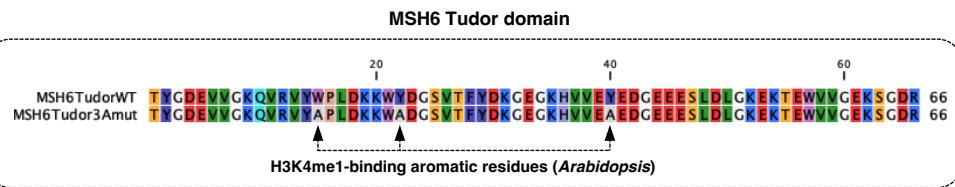
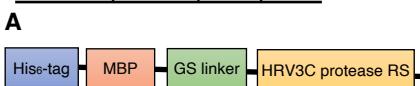
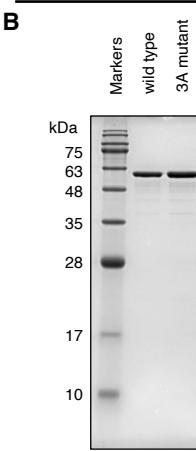


Figure 1. *Arabidopsis thaliana* and *Oryza sativa* PDS5C and MSH6 Tudor domains affinity to H3K4 methylation. **A-B**, AlphaFold2 structure of PDS5C and MSH6 protein from *A. thaliana* indicates Tudor domain is tethered to active domains **C**, ClustalW alignment of Tudor domains. Red arrows mark residues shown to be important for H3K4me1 specific binding, and yellow arrows mark residues important to interact with H3R2 (Niu et al. 2021) **D**, Aromatic cage and H3K4me1 in experimental structure of At_PDS5C (PDB: 7DE9) (Niu et al. 2021) highlighting in yellow the distances to the center of the three aromatic rings. **E**, % of models generated with an average distance < 6 Å. Docking benchmarking in relation to ITC experimental results from *A. thal* PDS5C. Average distance from the side chain nitrogen of H3K4 to the center of aromatic rings forming the aromatic cage (see B) of top 1,000 models (10%) based on Rosetta total score after docking H3 tail peptides in different methylation states to the Tudor domains of PDS5C and MSH6 from both *A. thaliana* and *O. sativa*. The dashed line marks the average distance in the experimental structure as a reference. **F-I**, FlexPepDock models of the H3K4me1 peptide in the aromatic cage of Tudor domains of At_PDS5C, At_MSH6, Os_PDS5C and OS_MSH6. Residues interacting with H3K4me1 are amplified in the left boxes. Residues interacting with H3R2 are highlighted in the right boxes.

Domain composition of purified proteins



SDS-PAGE with CBB staining



C

Arabidopsis MSH6 Tudor wildtype

Arabidopsis MSH6 Tudor w/ 3A Mutation

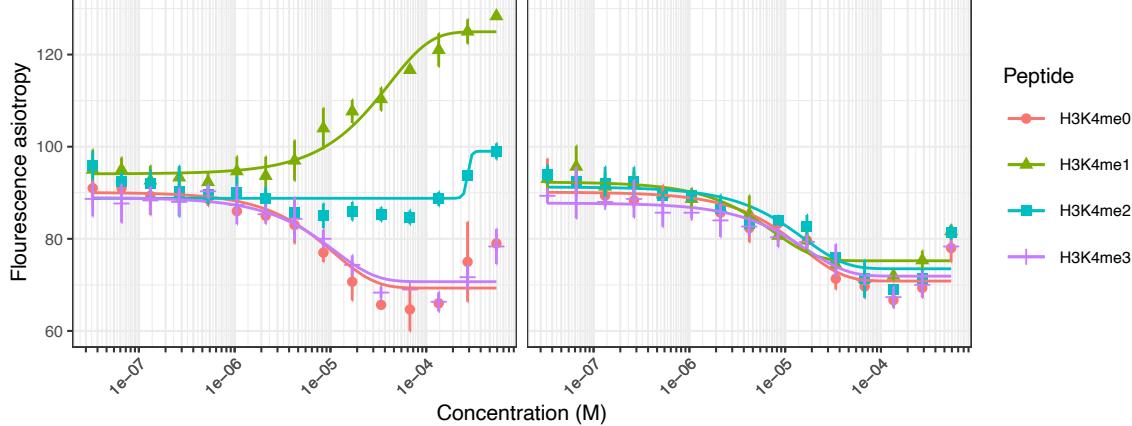


Figure 2. Experimental validation of *Arabidopsis* MSH6 Tudor domain affinity to H3K4 methylation. **A)** Domain and sequence of purified protein used in fluorescence polarization (FP). The 3A mutant protein only replaces aromatic residues that bind H3K4me1 from the wildtype Tudor domain with Alanine. **B)** SDS-PAGE confirming purity. **C)** FP results (3 replicates per experiment). Points represent mean and error bars standard deviations.

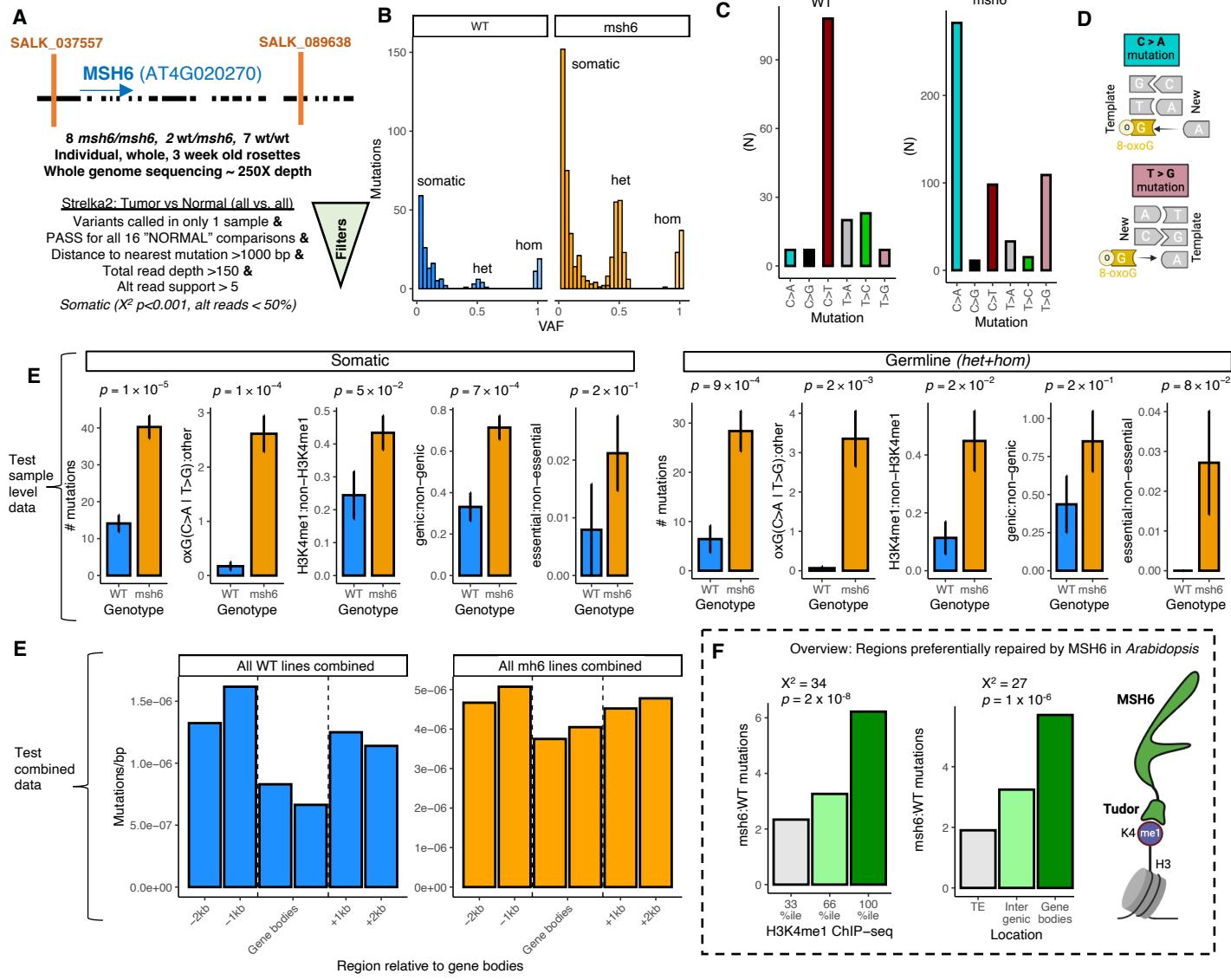


Figure 3. Somatic mutations in MSH6 knockout lines. **A**, Diagram representing MSH6 mutant SALK insertions used to phenotype mutation rates in MSH deficient lines. General pipeline for Somatic mutation experimental design and strict variant calling using Strelka2. **B**, Alternative allele frequency of mutations found in wt/wt+wt/msh6 and msh6/msh6 lines. Total number of mutations (n) are 127 and 322 respectively. **C**, mutation signature of wt/wt+wt/msh6 and msh6/msh6 with msh6/msh6 showing **D**, expected effects of unrepaired mismatches between adenine and oxidized guanine. **E**, Proportion of mutations distribution in gene bodies in wt/wt+wt/msh6 and msh6/msh6 lines. **F**, msh6/msh6 to wt/wt+wt/msh6 mutation ratio in different genomic regions grouped by H3K4me1 ChIP-seq.

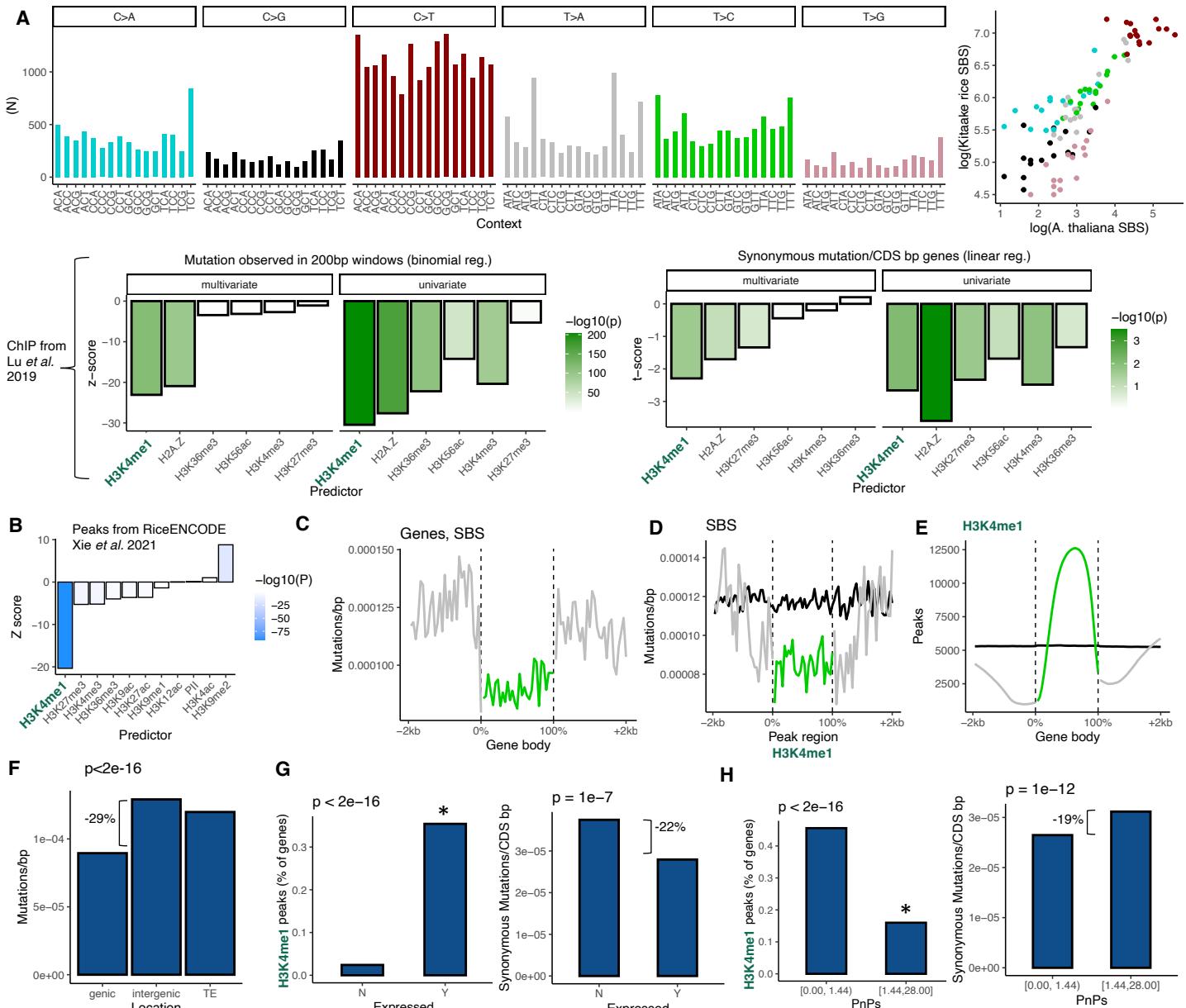


Figure 4. Spectra and distribution of germline SBS mutations in fast neutron rice. **A**, Frequency of mutations in trinucleotide contexts and correlation with *de novo* germline mutations in *A. thaliana* ($r>0.8$, $p<2\times10^{-16}$) (Weng et al. 2019, Monroe et al. 2022). **B**, Results from logistic regression modeling mutation probability in 100bp windows as a function of overlap with epigenomic marks around genes in rice. **C**, Mutation rates in relation to gene bodies in rice. **D**, Mutation rates in relation to H3K4me1 peak regions in rice. **E**, H3K4me1 peaks regions distribution in gene bodies in rice. **F**, Black line shows null expectation (mutations around randomized peaks). **F**, Mutation rates in genic, intergenic, and TE regions. **G**, H3K4me1 enrichment and gene body mutation rates in genes annotated as expressed (Y) and non-expressed (N) in rice. **H**, H3K4me1 enrichment and gene body mutation rates in genes subject variable degrees of purifying selection (low vs high Pn/Ps) in 3,010 natural accessions of rice.

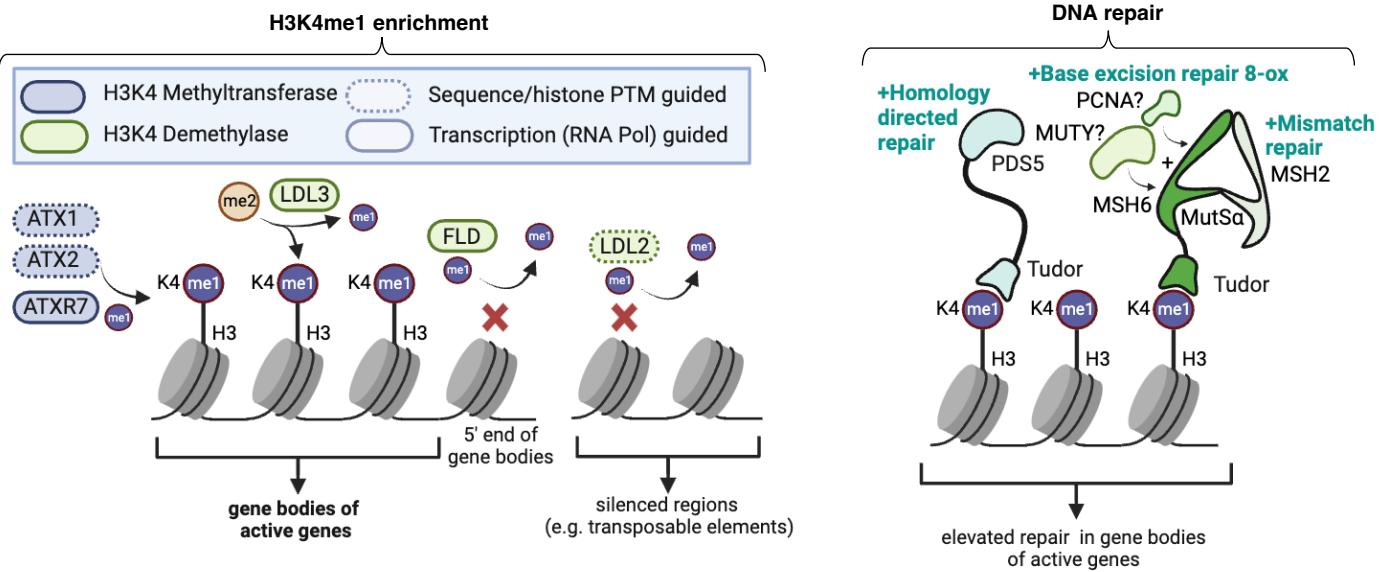


Figure 5. Emerging mechanistic model of H3K4me1-targeted DNA repair in plants. ATX1, ATX2, ATXR7 are responsible for the enrichment of H3K4me1 in the gene bodies of active genes (Oya et al. 2021). Tudor domain of PDS5C binds H3K4me1 and facilitates homology directed repair (HDR) (Niu et al. 2021). Other repair genes including PDS5 paralogs and MSH6 (responsible for mismatch repair – MMR) also contain functionally similar Tudor domains, with H3K4me1 targeting potential.

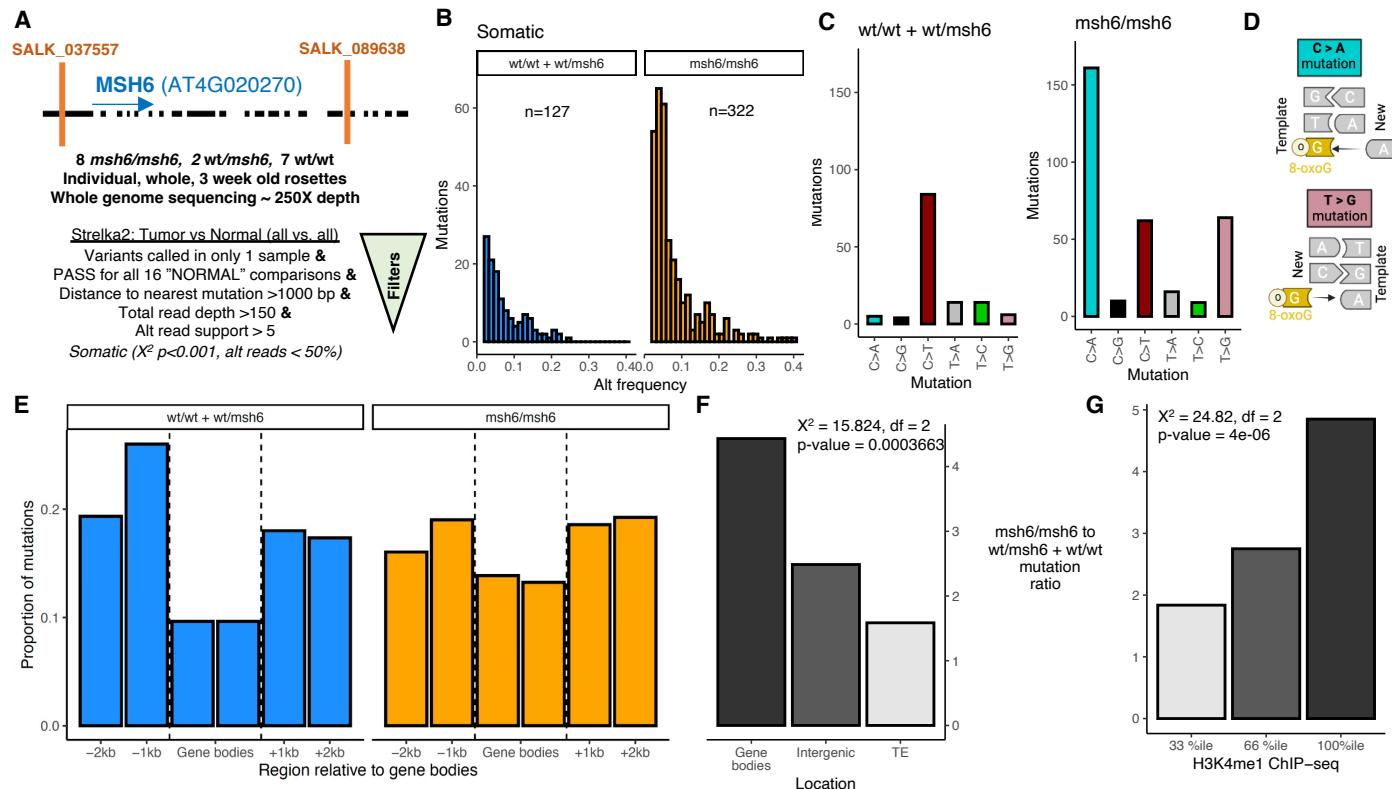
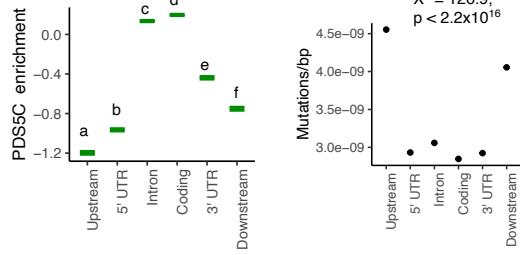
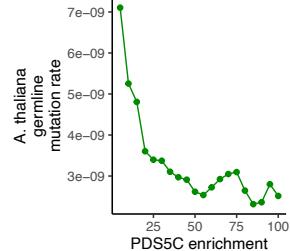
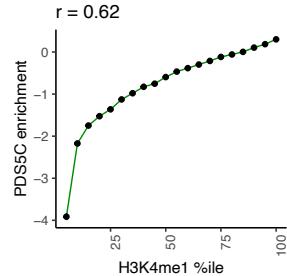
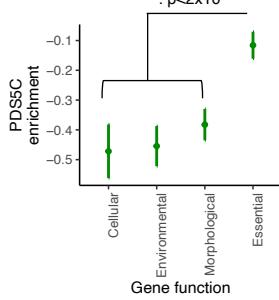
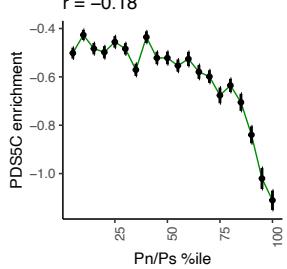


Figure 3. Somatic mutations in MSH6 knockout lines. **A**, Diagram representing MSH6 mutant SALK insertions used to phenotype mutation rates in MSH deficient lines. General pipeline for Somatic mutation experimental design and strict variant calling using Strelka2. **B**, Alternative allele frequency of mutations found in wt/wt+wt/msh6 and msh6/msh6 lines. Total number of mutations (n) are 127 and 322 respectively. **C**, mutation signature of wt/wt+wt/msh6 and msh6/msh6 with msh6/msh6 showing **D**, expected effects of unrepaired mismatches between adenine and oxidized guanine. **E**, Proportion of mutations distribution in gene bodies in wt/wt+wt/msh6 and msh6/msh6 lines. **F**, msh6/msh6 to wt/wt+wt/msh6 mutation ratio in different genomic features and **G**, in genome regions grouped by H3K4me1 ChIP-seq.

A**B****C****D****E**

I think the one that I left out is the less relevant to the analysis, but I don't know... I just took it out to make it symmetrical.

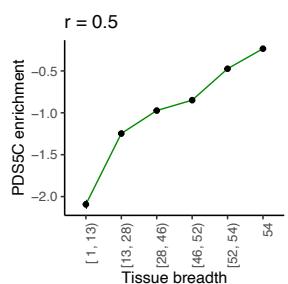


Figure 2. PDS5C targets H3K4me1 and is associated with lower mutation rates. **A**, ChIP-seq analysis of PDS5C in *A. thaliana*, cohesion cofactor involved in HDR (Niu et al. 2021). Letters represent TukeyHSD contrast at $p=0$. Upstream/downstream = 1000bp. **B**, Features marked by PDS5C experience reduced germline mutation rates in *A. thaliana*. **C**, Gene bodies of genes with elevated H3K4me1 have greater levels of PDS5C, as do **D**, essential genes, and **E**, genes under stronger purifying selection. Pn/Ps = Non-synonymous /synonymous polymorphisms in *A. thaliana* populations. Mean + SE for percentiles shown. Pearson correlation (r) reflects entire dataset. $P<2\times 10^{-16}$ for all.

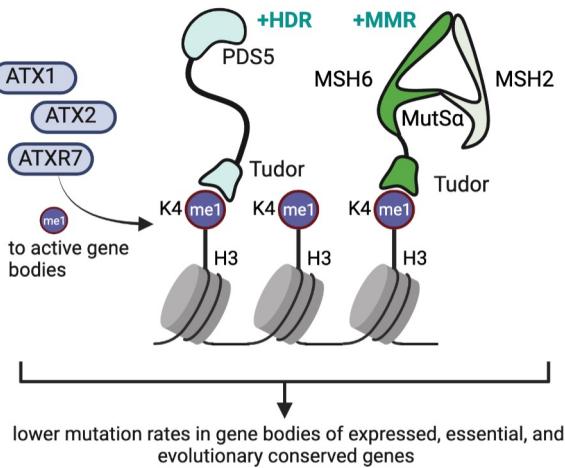


Figure 5. Emerging mechanistic model of H3K4me1-targeted DNA repair in plants. ATX1, ATX2, ATXR7 are responsible for the enrichment of H3K4me1 in the gene bodies of active genes (Oya et al. 2021). Tudor domain of PDS5C binds H3K4me1 and facilitates homology directed repair (HDR) (Niu et al. 2021). Other repair genes including PDS5 paralogs and MSH6 (responsible for mismatch repair – MMR) also contain functionally similar Tudor domains, with H3K4me1 targeting potential.

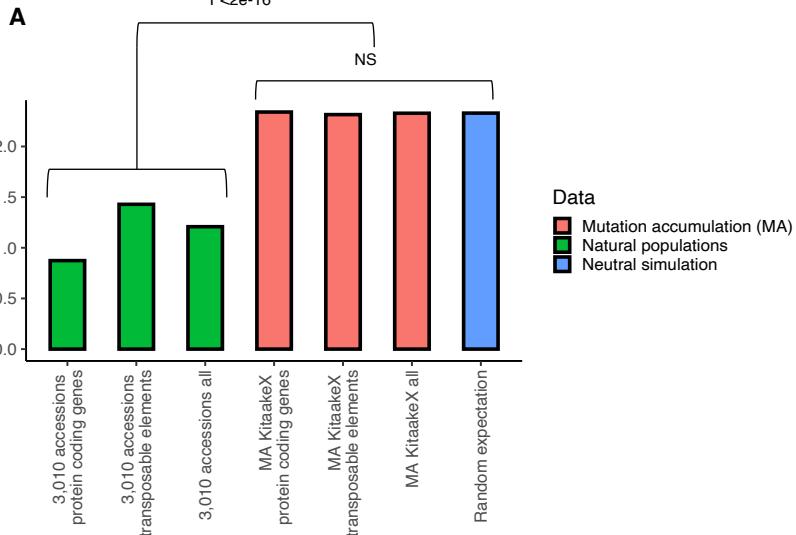


Figure S1. Non-synonymous to synonymous ratios. The non-synonymous to synonymous ratio was significantly lower ($p<2\times 10^{-16}$) in natural populations compared to mutation accumulation lines and neutral expectation. The non-synonymous to synonymous ratio was similar in mutation accumulation lines to neutral expectation.

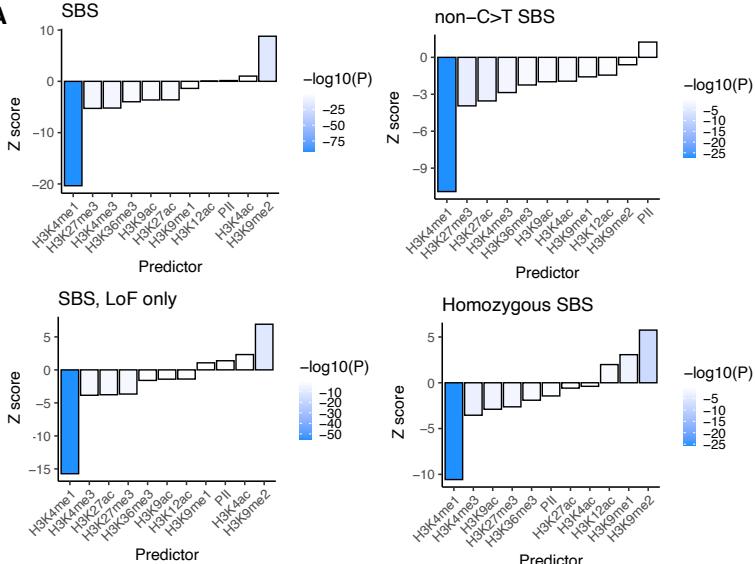
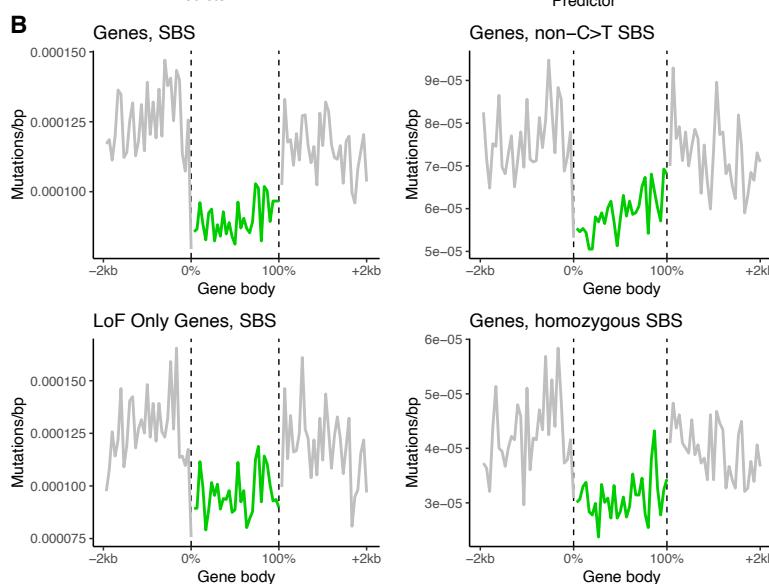
A**B**

Figure S2. Relationships between epigenomic features, mutation and gene bodies

A, Results from logistic regression modeling mutation probability in 100bp windows as a function of overlap with epigenomic marks around genes in rice. Four panels reflect various subsets of data to test for effects of cytosine deamination in transposable elements as alternative explanation, selection, or mapping.

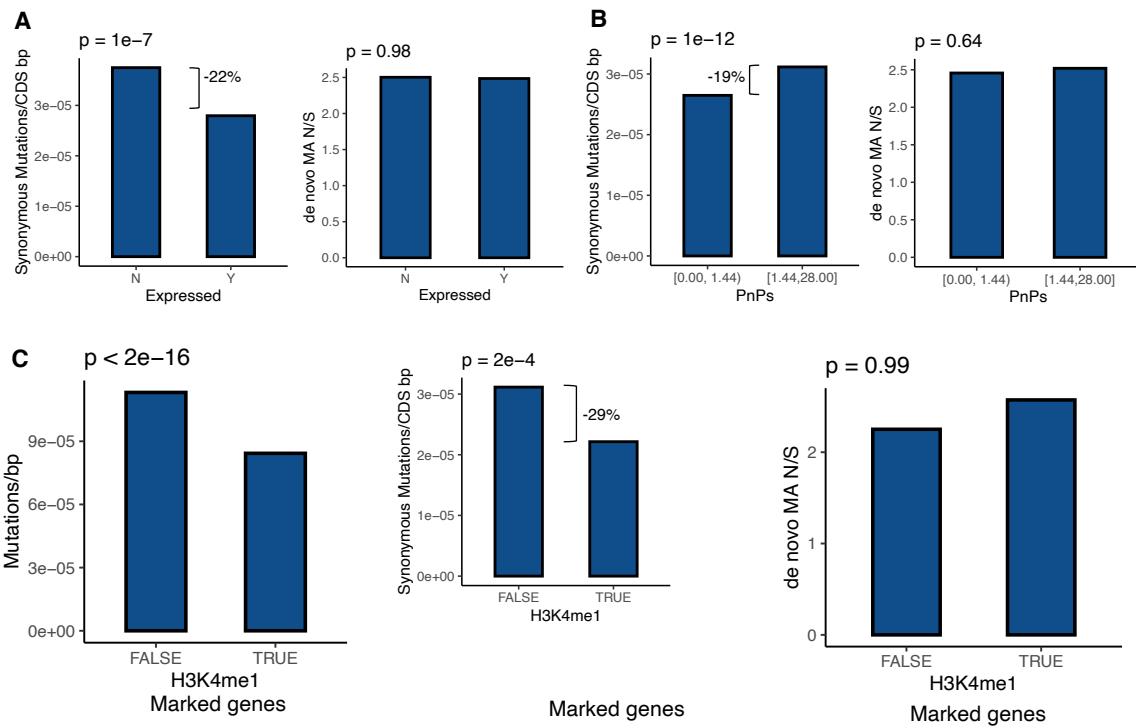
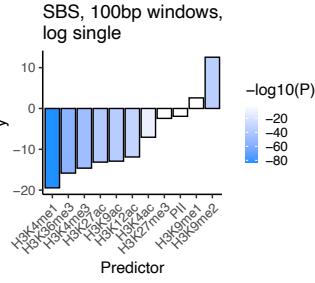
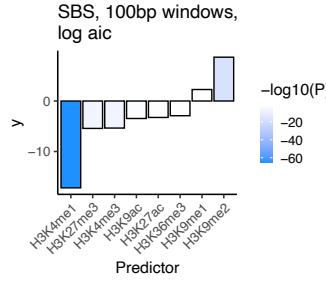


Figure S3. Expressed and conserved genes and H3K4me1 marked genes have lower mutation rates in rice A, synonymous mutation rates, non-synonymous to synonymous *de novo* mutation ratio in mutation accumulation lines (MA N/S) in expressed vs non-expressed genes and B, in genes subject variable degrees of purifying selection (low vs high Pn/Ps) in 3,010 natural accessions of rice. C, mutation rates, non-synonymous to synonymous *de novo* mutations in mutation accumulation lines (MA N/S) in H3K4me1 non-marked genes (FALSE) and marked genes (TRUE).

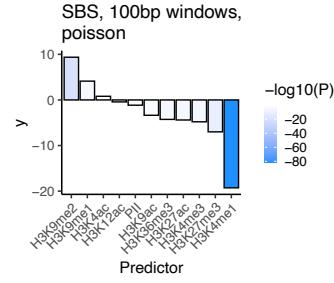
**Single variable
binomial (logistic)
regression**



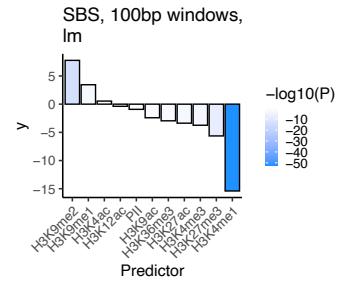
**Multiple regression,
binomial with AIC
model selection**



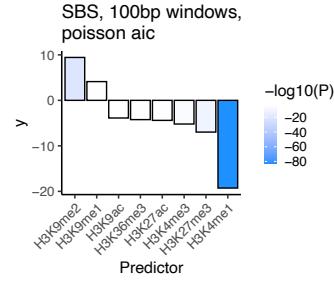
**Multiple regression,
Poisson**



**Multiple regression,
linear**



**Multiple regression,
Poisson, with AIC
model selection**



**Multiple regression,
linear, with AIC
model selection**

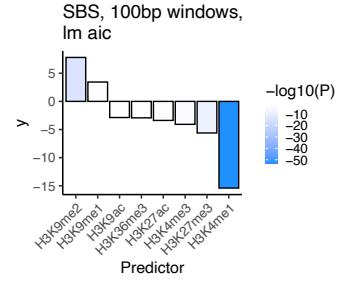


Figure S4. Supplementary models: Relationships between epigenomic features, mutation and gene bodies.

I simplified and only show the extra modelling approaches for all SBS (instead of all the subsets like non C.T)

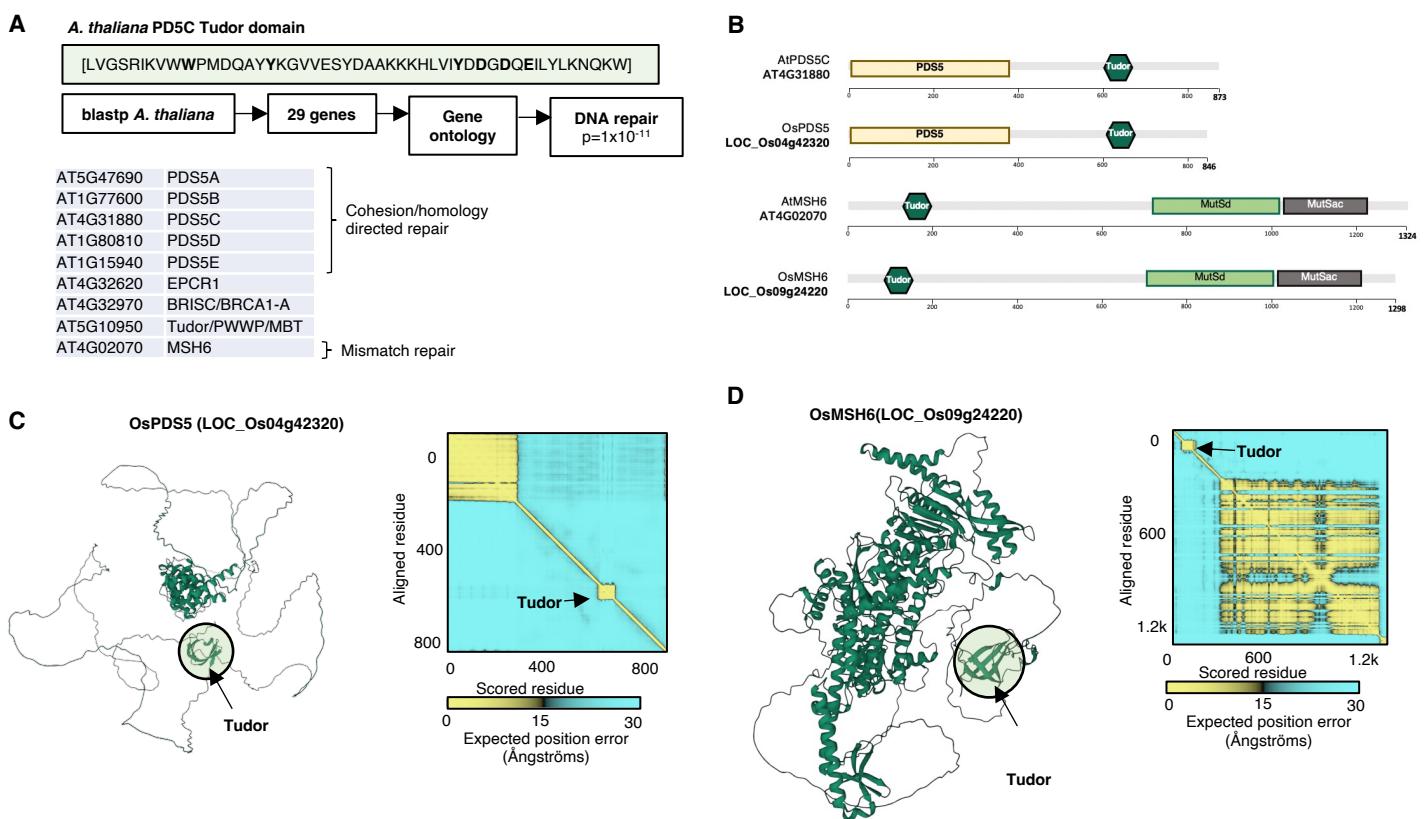
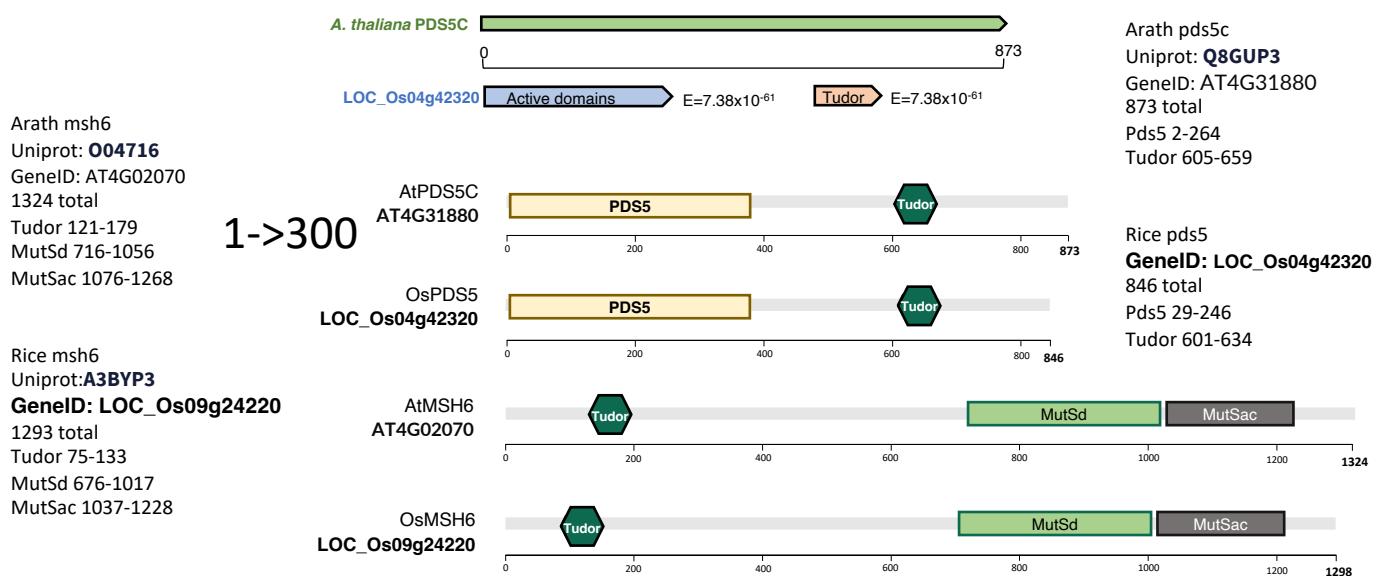


Figure S5. Multiple repair pathways with H3K4me1 targeting potential via Tudor domains. **A**, Results from blastp of PCDS5C Tudor domain, which has confirmed H3K4me1 binding specificity (Niu et al. 2021) identified 29 genes. Nine of these are annotated as having DNA repair function (significant enrichment by gene ontology analysis). **B**, Domain prediction of *A. thaliana* and *O. sativa* putative orthologs of PDS5C and MSH6 genes. **C-D**, Structure of PDS5C and MSH6 protein from AlphaFold2 indicates Tudor domain is tethered to active domains structure in putative *O. sativa* orthologs.





- █ *A. thaliana* PDS5C Tudor Domain
(PDB: 7DE9)
- █ *O. sativa* PDS5C Tudor Domain
(AF model)
- █ *A. thaliana* MSH6 Tudor domain
(AF model)
- █ *O. sativa* MSH6 Tudor domain
(AF model)

Figure S6. AlphaFold predictions of Tudor domains. AlphaFold models of *A. thaliana* and *O. sativa* MSH6 and *O. sativa* PDS5C Tudor are shown superimposed onto the experimental structure of *A. thaliana* PDS5C Tudor domain (PDB:7DE9) (Niu et al. 2021).

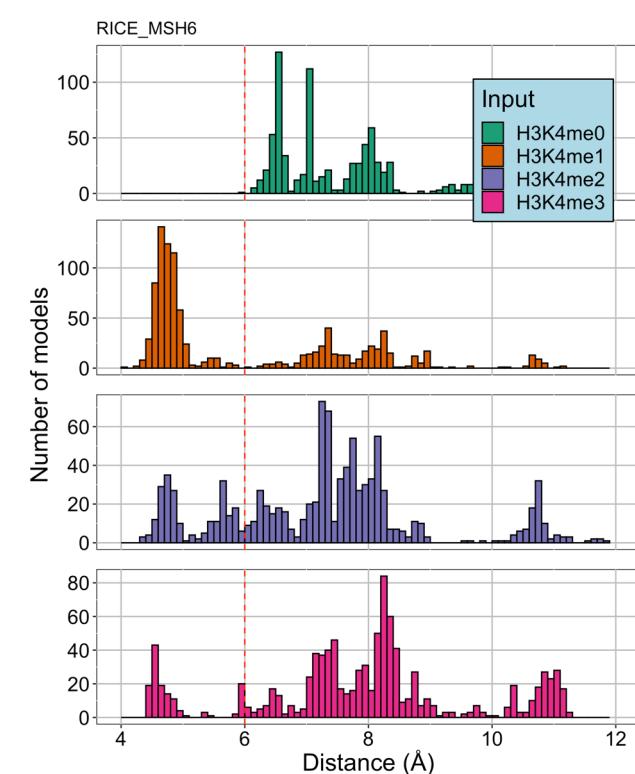
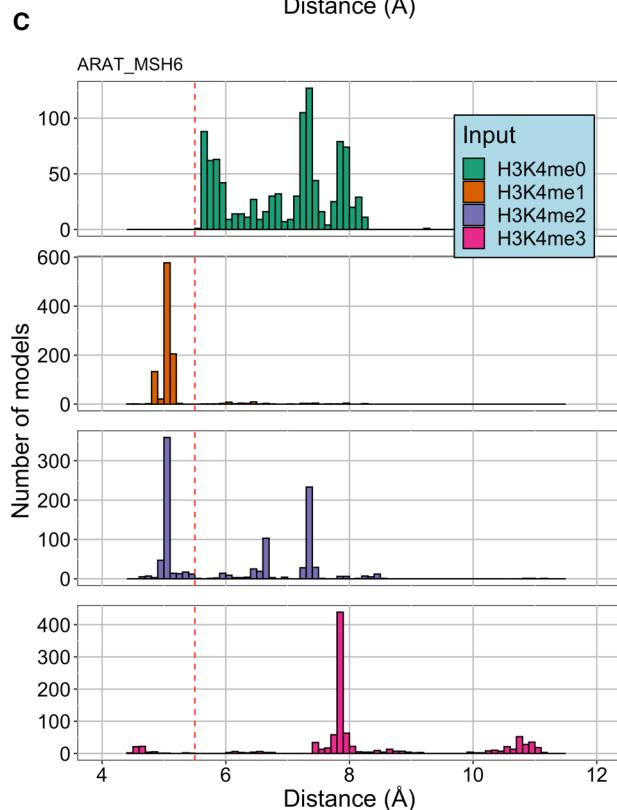
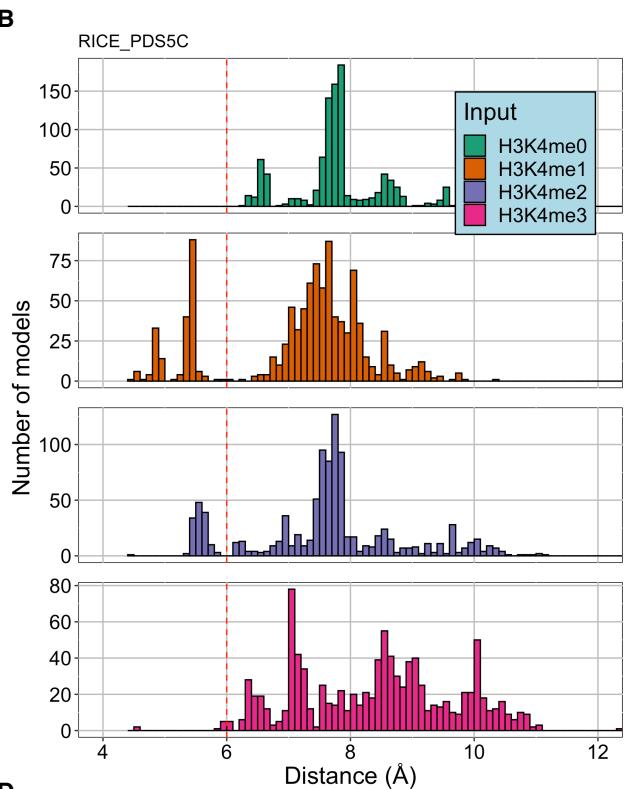
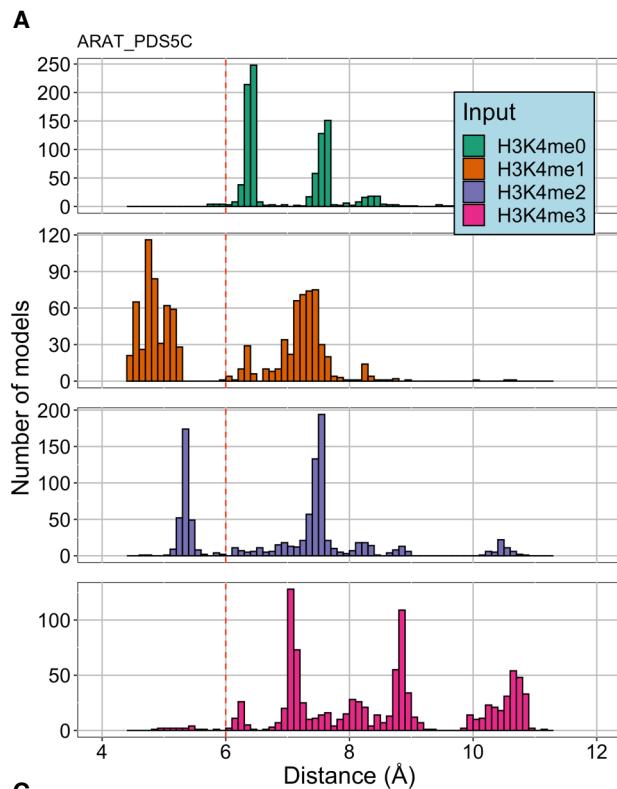


Figure S7. The average distance (in \AA) distribution to the aromatic cage in FlexPepDock models of the different H3K4 marks tested, for; **A**, *A. thaliana* PDS5C Tudor experimental structure (PDB:7de9) **B**, *O. sativa* PDS5C Tudor Alphafold model **C**, *A. thaliana* MSH6 Tudor Alphafold model **D**, *O. sativa* MSH6 Tudor Alphafold model. **A-D**, The red line marks the threshold of 6 \AA (defining the H3K4 inside the aromatic cage).

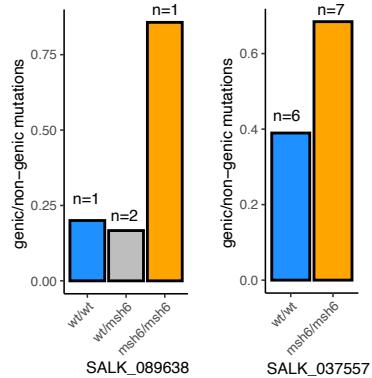
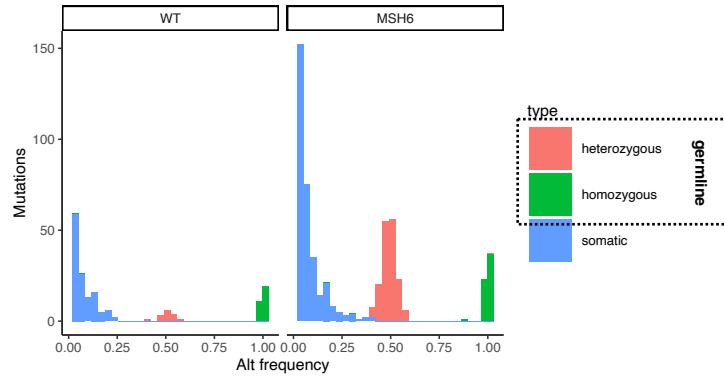
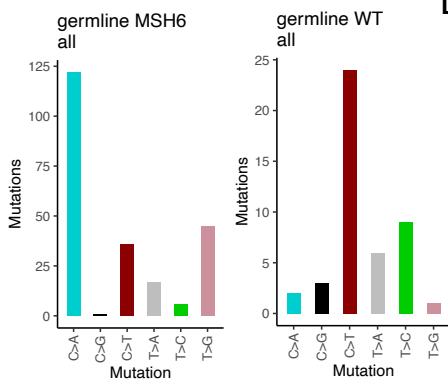
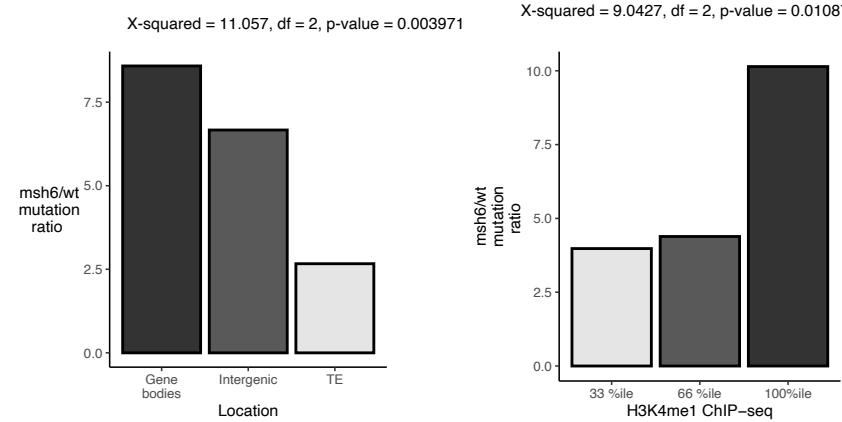
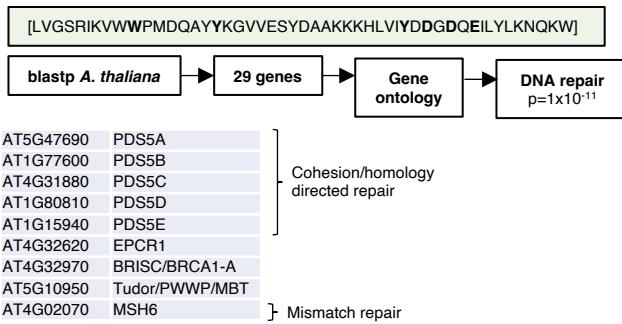
A**B****C****D**

Figure S7. **A**, genic/non-genic mutations in MSH6 knockout lines SALK_089638 and SALK_037557. **B**, frequency of alternative alleles distribution detected by Strelka2 variant caller. **C**, Germline mutations signature in MSH6 knockout lines and WT. **D**, msh6/wt mutation ratio in different genomic features and in different H3K4me1 ChIP-seq %ile for *de novo* germline mutations.

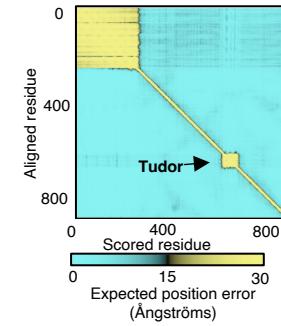
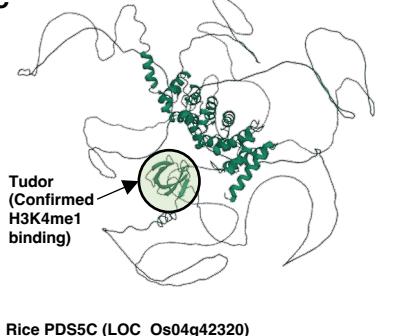
A. thaliana PD5C Tudor domain



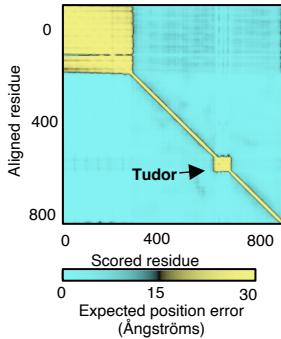
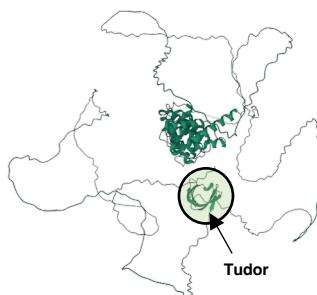
B



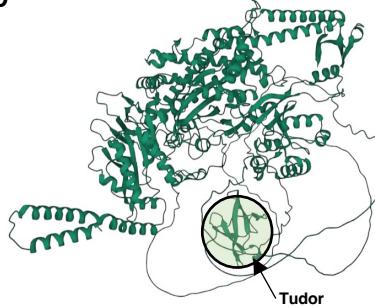
C A. thaliana PDS5C



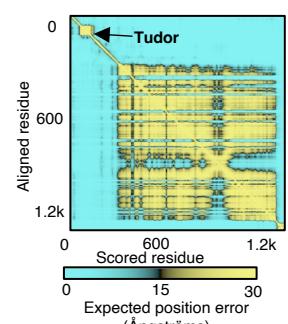
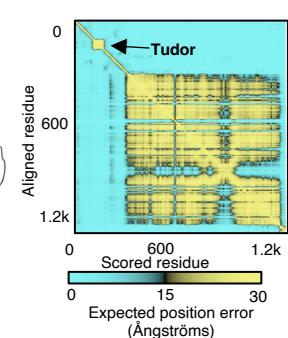
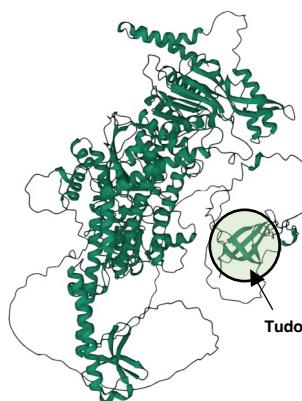
Rice PDS5C (LOC_Os04g42320)



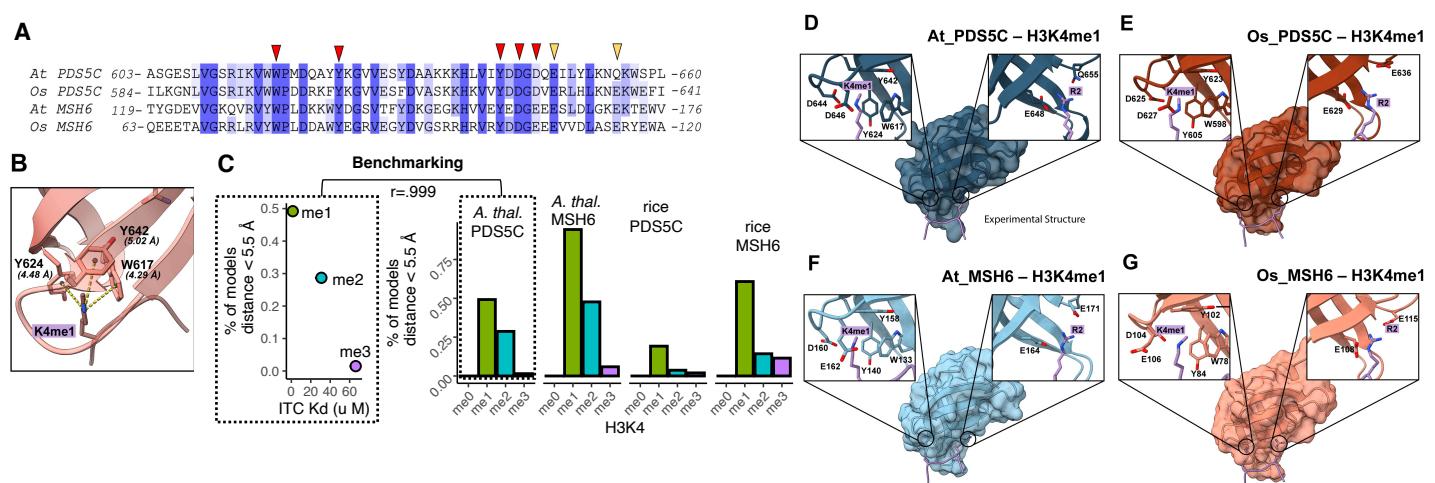
D A. thaliana Mismatch repair MSH6



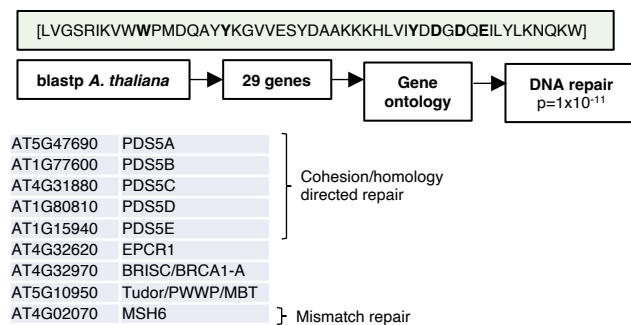
Rice Mismatch repair MSH6



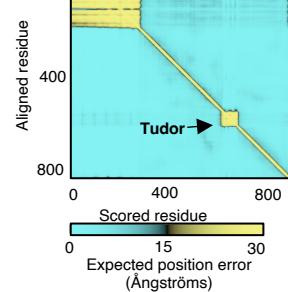
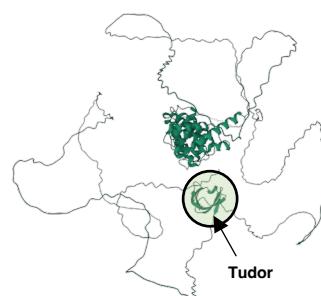
Alternative Figure S4. Multiple repair pathways with H3K4me1 targeting potential via Tudor domains. **A**, Results from blastp of PCDS5C Tudor domain, which has confirmed H3K4me1 binding specificity (Niu et al. 2021) identified 29 genes. Nine of these are annotated as having DNA repair function (significant enrichment by gene ontology analysis). **B**, Protein domain structure and comparison of *A. thaliana* PDS5C and putative rice ortholog LOC_Os04g43220, and *A. thaliana* MSH6 and putative rice ortholog LOC_Os09g24220. **C-D**, Structure of PDS5C and MSH6 protein from AlphaFold2 indicates Tudor domain is tethered to active domains and similar structure observed in putative rice orthologs.



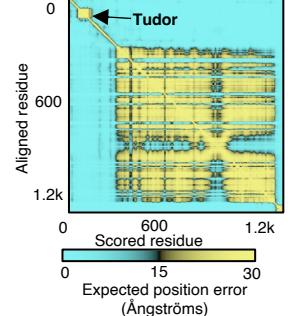
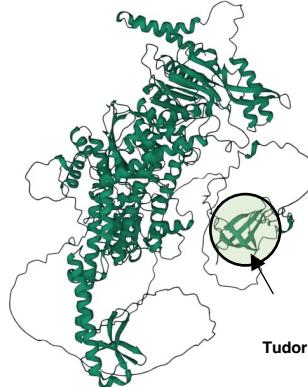
Alternative Figure 3. *Arabidopsis thaliana* and *Oryza sativa* PDS5C and MSH6 Tudor domains affinity to H3K4 methylation. **A**, ClustalW alignment of Tudor domains. Red arrows mark residues shown to be important for H3K4me1 specific binding, and yellow arrows mark residues important to interact with H3R2 (Niu et al. 2021) **B**, Aromatic cage and H3K4me1in experimental structure of At_PDS5C (PDB: 7DE9) (Niu et al. 2021) highlighting in yellow the distances to the center of the three aromatic rings. **C**, % of models generated with an average distance < 6 Å. Docking benchmarking in relation to ITC experimental results from *A. thal.* PDS5C. Average distance from the side chain nitrogen of H3K4 to the center of aromatic rings forming the aromatic cage (see B) of top 1,000 models (10%) based on Rosetta total score after docking H3 tail peptides in different methylation states to the Tudor domains of PDS5C and MSH6 from both *A. thaliana* and *O. sativa*. The dashed line marks the average distance in the experimental structure as a reference. **D-G**, FlexPepDock models of the H3K4me1 peptide in the aromatic cage of Tudor domains of At_PDS5C, At_MSH6, Os_PDS5C and OS_MSH6. Residues interacting with H3K4me1 are amplified in the left boxes. Residues interacting with H3R2 are highlighted in the right boxes.

A *A. thaliana* PD5C Tudor domain**B**

OsPDS5 (LOC_Os04g42320)

**C**

OsMSH6(LOC_Os09g24220)



Alternative Figure S5. Multiple repair pathways with H3K4me1 targeting potential via Tudor domains. **A**, Results from blastp of PCDS5C Tudor domain, which has confirmed H3K4me1 binding specificity (Niu et al. 2021) identified 29 genes. Nine of these are annotated as having DNA repair function (significant enrichment by gene ontology analysis). **B-C**, Structure of PDS5C and MSH6 protein from Alphafold2 indicates Tudor domain is tethered to active domains structure in putative *O. sativa* orthologs.

SALK_037557

SALK_089638

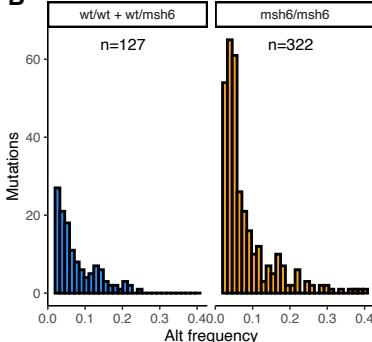
A MSH6 (AT4G020270)

8 msh6/msh6, 2 wt/msh6, 7 wt/wt
Individual, whole, 3 week old rosettes
Whole genome sequencing ~ 250X depth

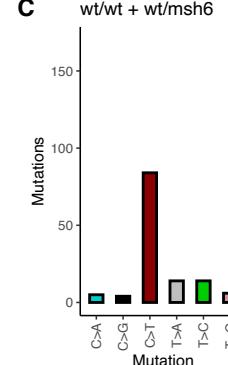
Strelka2: Tumor vs Normal (all vs. all)
PASS for all 16 "NORMAL" comparisons &
Distance to nearest mutation >1000 bp &
Total read depth >150 &
Alt read support > 5
Somatic (χ^2 p<0.001, alt reads < 50%)



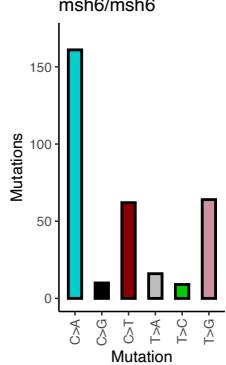
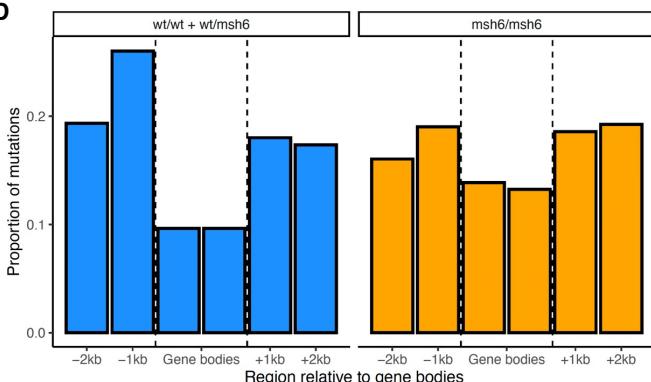
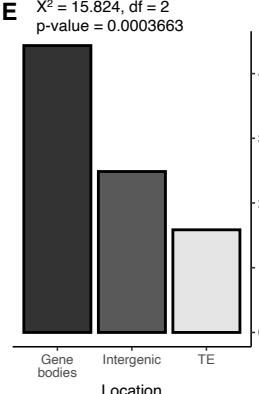
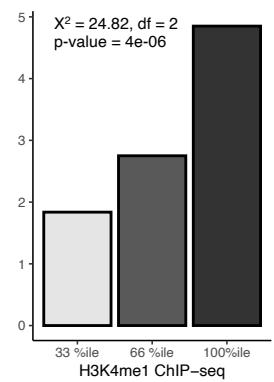
Somatic

**C**

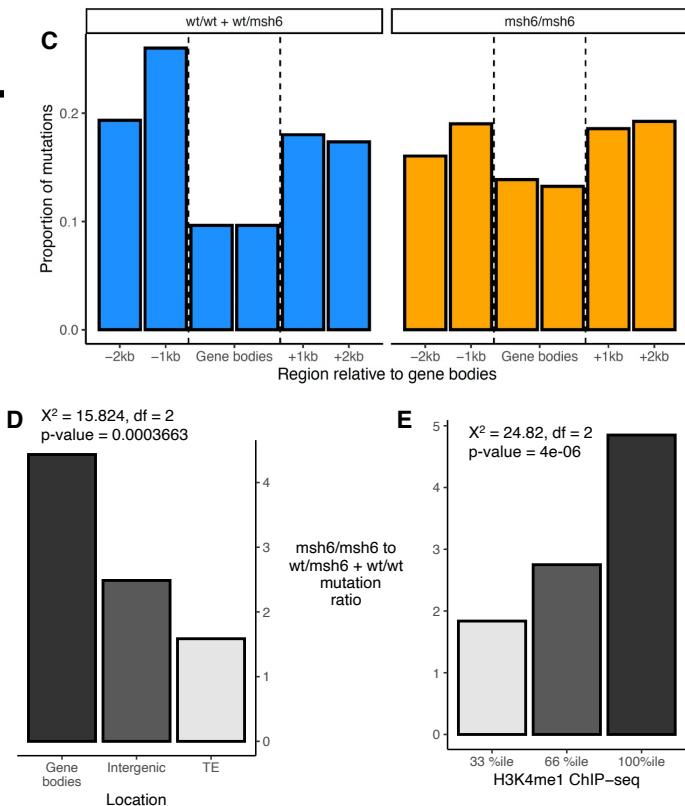
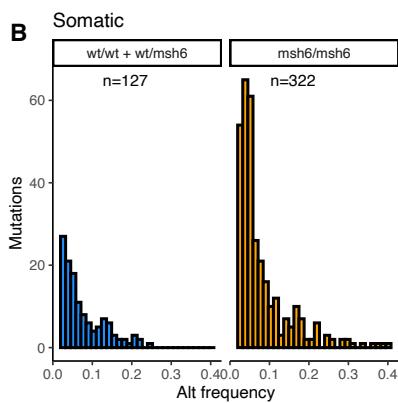
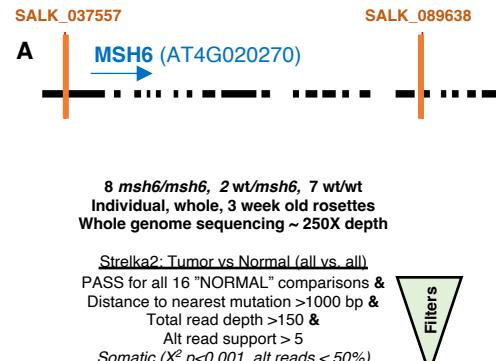
wt/wt + wt/msh6



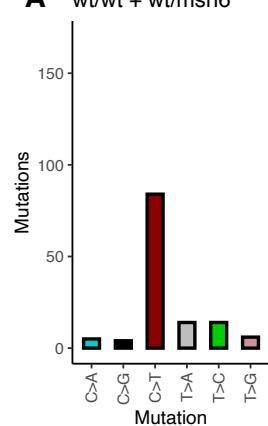
msh6/msh6

**D**
 $\chi^2 = 15.824, df = 2$
p-value = 0.0003663
**F**

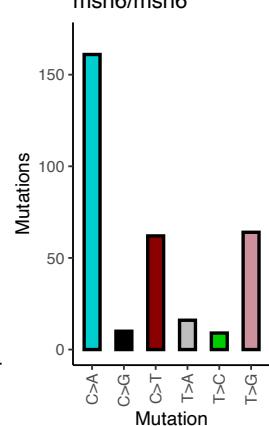
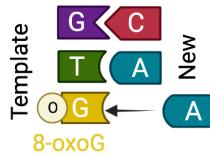
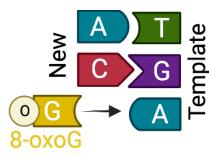
Alternative Figure 4. Somatic mutations in MSH6 knockout lines. **A**, Diagram representing MSH6 mutant SALK insertions used to phenotype mutation rates in MSH deficient lines. General pipeline for Somatic mutation experimental design and strict variant calling using Strelka2. **B**, Alternative allele frequency of mutations found in wt/wt+wt/msh6 and msh6/msh6 lines. Total number of mutations (n) are 127 and 322 respectively. **C**, mutation signature of wt/wt+wt/msh6 and msh6/msh6 showing expected effects of unrepaired mismatches between adenine and oxidized guanine. **D**, Proportion of mutations distribution in gene bodies in wt/wt+wt/msh6 and msh6/msh6 lines. **E**, msh6/msh6 to wt/wt+wt/msh6 mutation ratio in different genomic features and **F**, in genome regions grouped by H3K4me1 ChIP-seq.



Alternative Figure 4. Somatic mutations in MSH6 knockout lines. **A**, Diagram representing MSH6 mutant SALK insertions used to phenotype mutation rates in MSH deficient lines. General pipeline for Somatic mutation experimental design and strict variant calling using Strelka2. **B**, Alternative allele frequency of mutations found in wt/wt+wt/msh6 and msh6/msh6 lines. Total number of mutations (n) are 127 and 322 respectively. **C**, Proportion of mutations distribution in gene bodies in wt/wt+wt/msh6 and msh6/msh6 lines. **D**, msh6/msh6 to wt/wt+wt/msh6 mutation ratio in different genomic features and **E**, in genome regions grouped by H3K4me1 ChIP-seq.

A wt/wt + wt/msh6

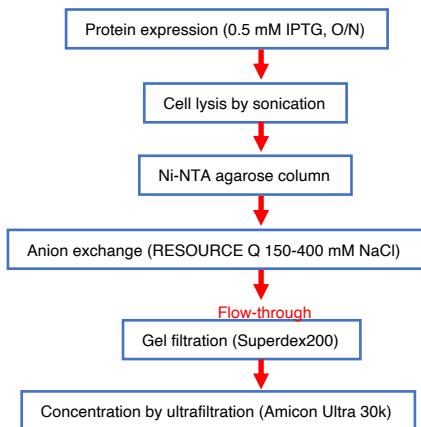
msh6/msh6

**B****C to A**
mutation**T to G**
mutation

Alternative Figure 4. Somatic mutations in MSH6 knockout lines. A, mutation signature of wt/wt+wt/msh6 and msh6/msh6 with msh6/msh6 showing B, expected effects of unrepaired mismatches between adenine and oxidized guanine.

Purification scheme

Cell: BL21(DE3) codonplusRIL
Scale: 5 L culture with LB medium



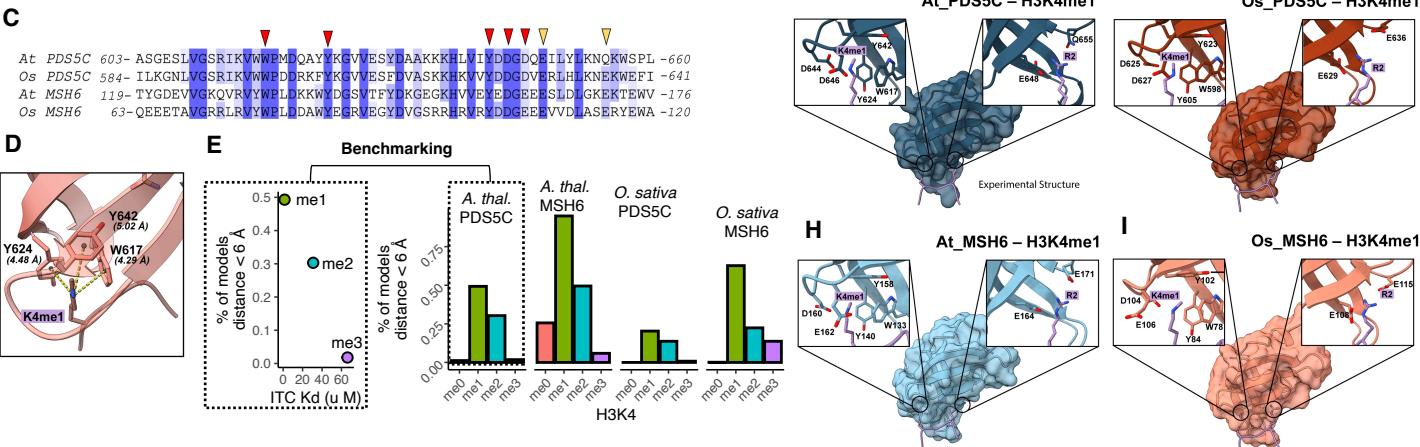
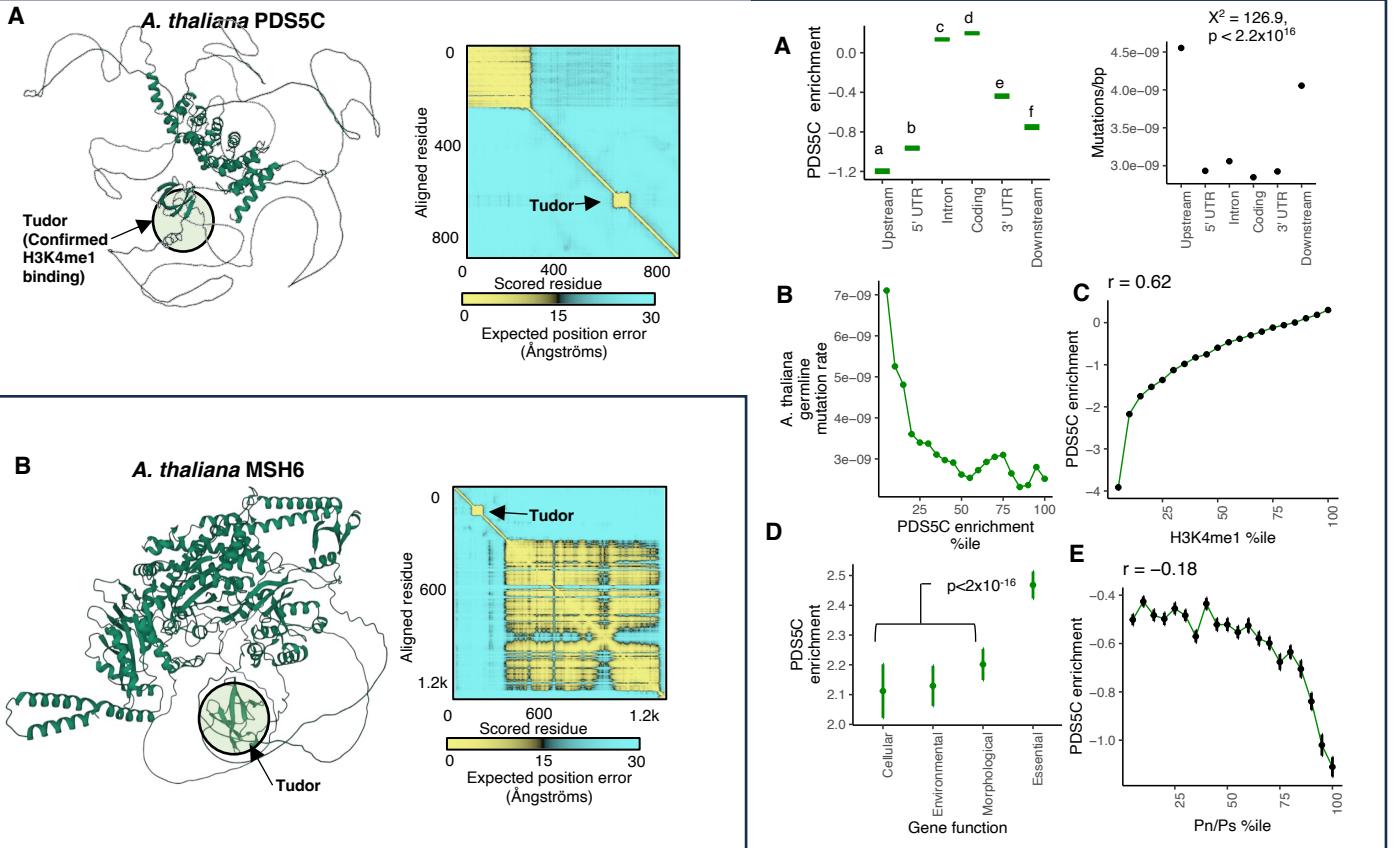


Figure 3. *Arabidopsis thaliana* and *Oryza sativa* PDS5C and MSH6 Tudor domains affinity to H3K4 methylation. **A-B**, AlphaFold2 structure of PDS5C and MSH6 protein from *A. thaliana* indicates Tudor domain is tethered to active domains **C**, ClustalW alignment of Tudor domains. Red arrows mark residues shown to be important for H3K4me1 specific binding, and yellow arrows mark residues important to interact with H3R2 (Niu et al. 2021) **D**, Aromatic cage and H3K4me1in experimental structure of At_PDS5C (PDB: 7DE9) (Niu et al. 2021) highlighting in yellow the distances to the center of the three aromatic rings. **E**, % of models generated with an average distance < 6 Å. Docking benchmarking in relation to ITC experimental results from *A. thal* PDS5C. Average distance from the side chain nitrogen of H3 tail peptides in different methylation states to the Tudor domains of PDS5C and MSH6 from both *A. thaliana* and *O. sativa*. The dashed line marks the average distance in the experimental structure as a reference. **F-I**, FlexPepDock models of the H3K4me1 peptide in the aromatic cage of Tudor domains of At_PDS5C, At_MSH6, Os_PDS5C and OS_MSH6. Residues interacting with H3K4me1 are amplified in the left boxes. Residues interacting with H3R2 are highlighted in the right boxes.