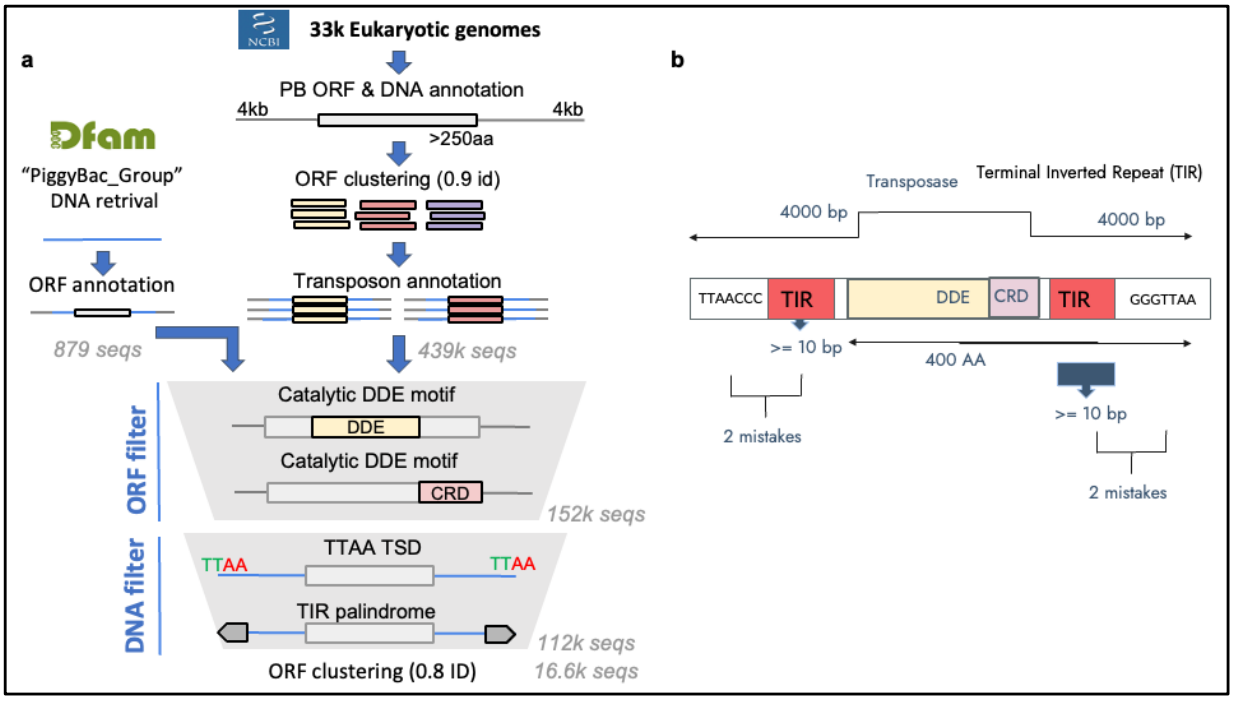
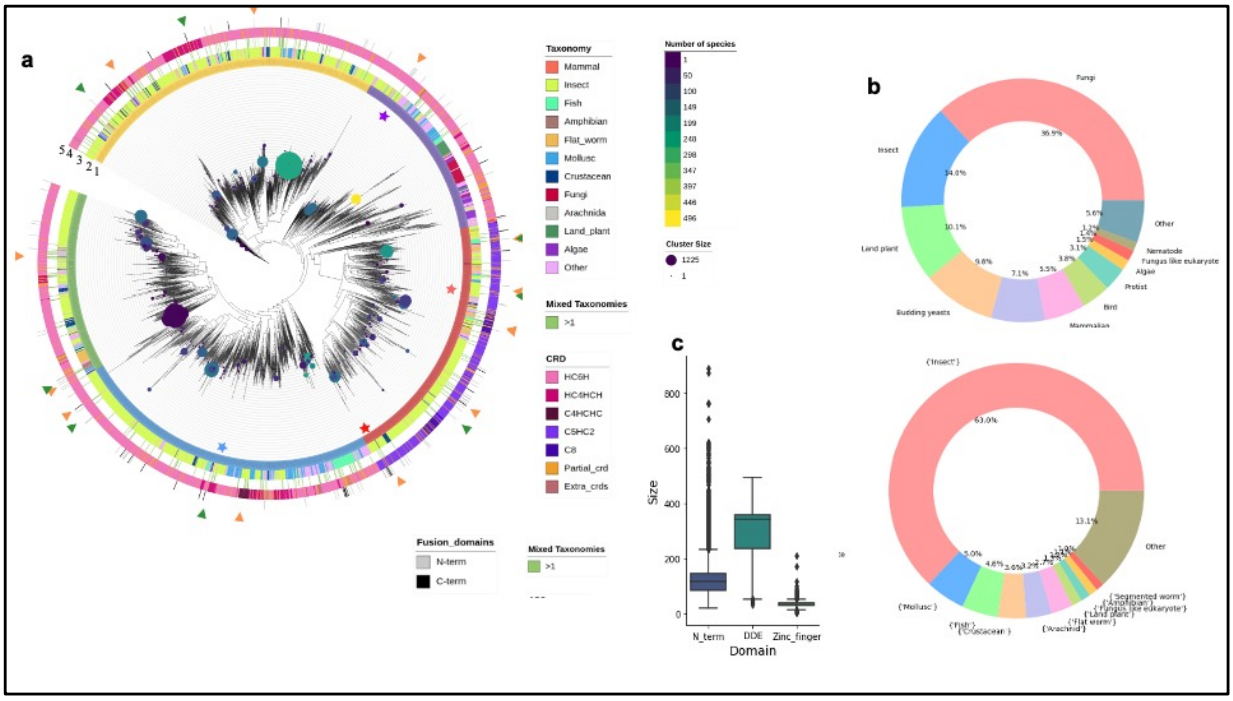
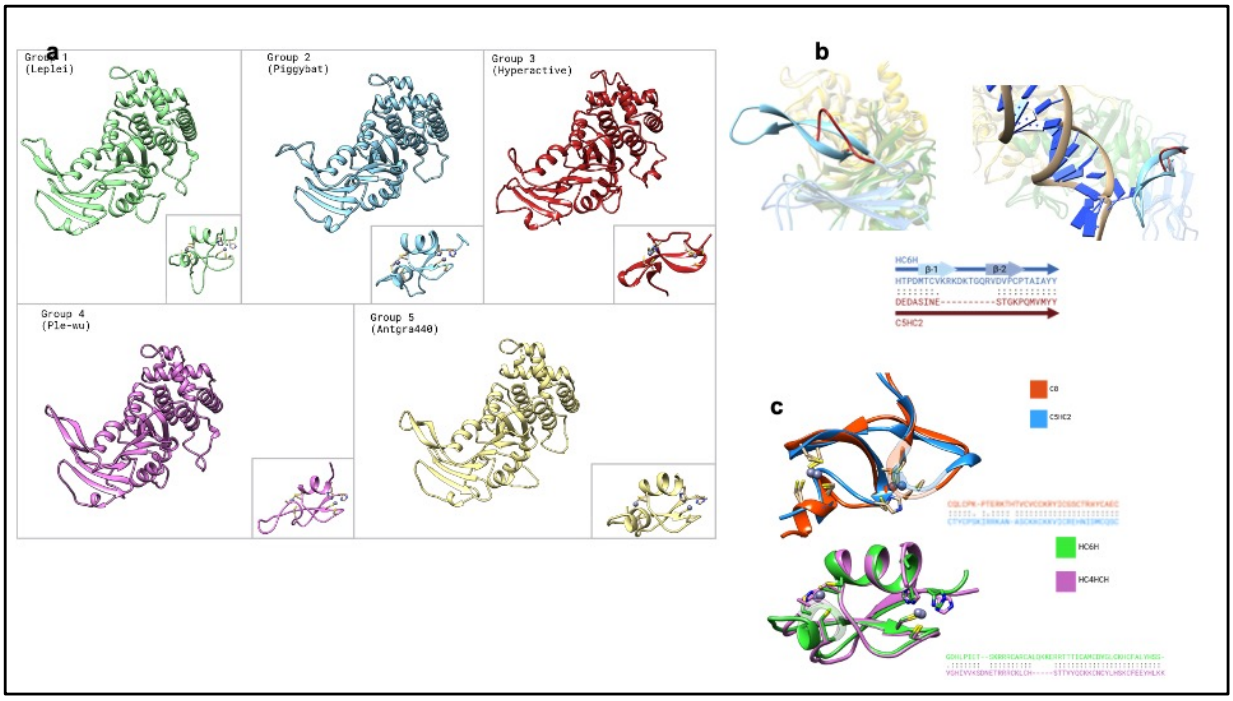


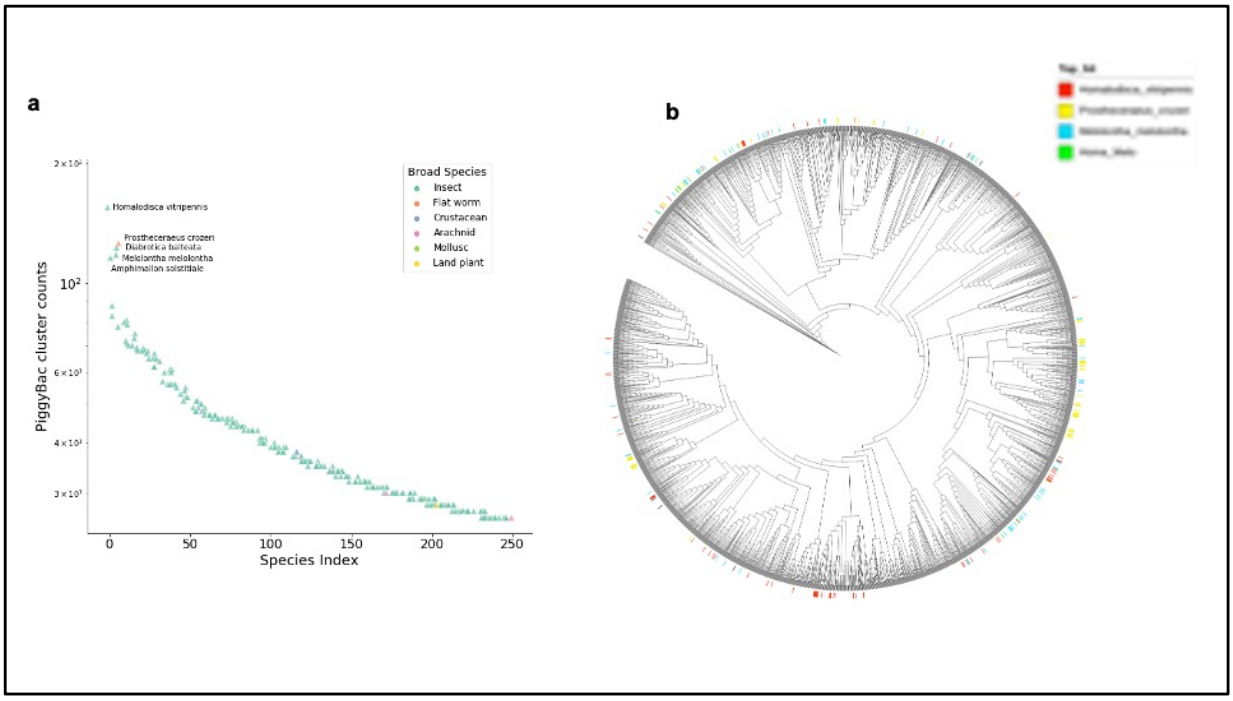
Supplementary Figure 1 | PiggyBac bioprospecting pipeline. a) Detailed of PiggyBac bioprospecting pipeline. Eukaryotic genomes were downloaded from ncbi and PiggyBac transposase ORFs were identified with HMM based search. ORFs were clustered at 0.9 id and DNA sequences spanning the ORF and 4kb at each end were fetched and aligned to anotate the transposase. In paralel, PiggyBac transposases from Dfam were retrieved. Putative transposon sequences were then filtered. ORFs were filtered to contain a catalytic DDE(D) triad, and corresponding cysteines in the ZnF containing cysteine rich domain (CRD). DNA ends were filtered to contain TTAA target strand duplication (TSD), and at least one terminally inverted palindrome. After applying the ORF and the DNA filters, ORF clustering at 0.8 was performed to obtain the final dataset. TIR and TSD DNA filtering removed 5,000 sequences (~2%) suggesting that feature conservation at the aminoacid level is a strong indicator of activity. **b)** Transposon DNA architecture and filtering details.



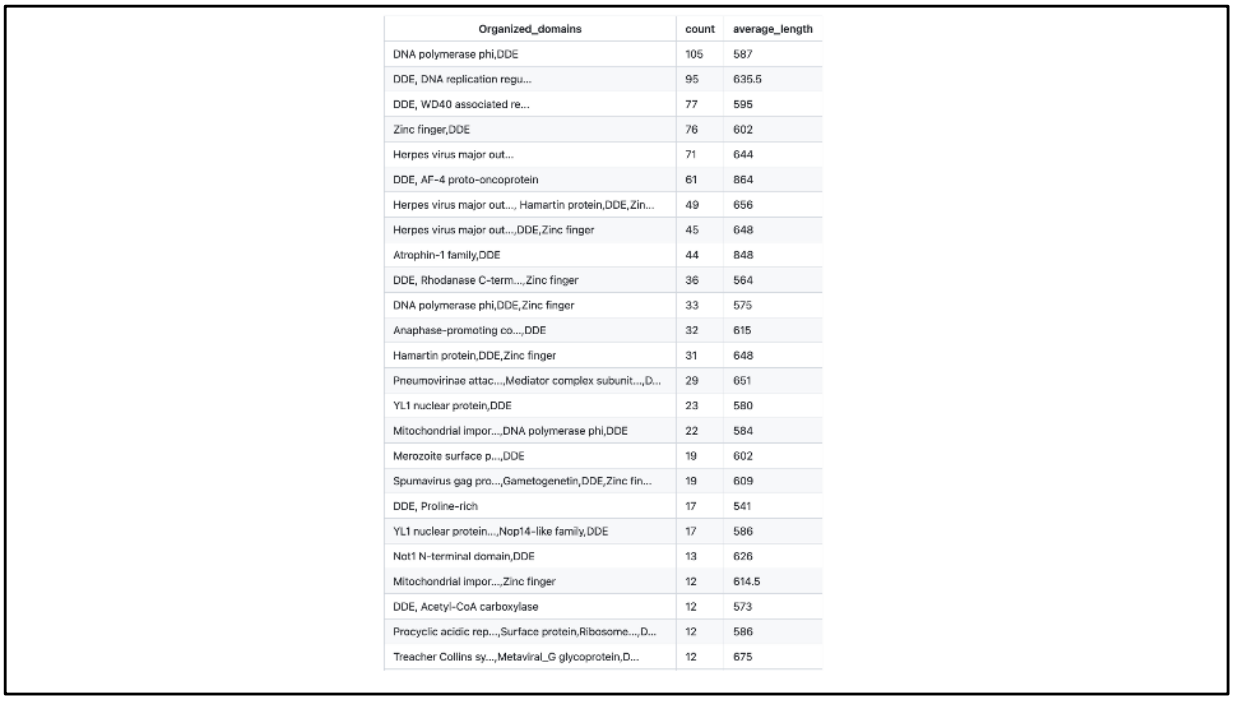
Supplementary Figure 2 | Full piggyBac phylogenetic tree with legends. a) PiggyBac tree from figure 1 showing all the ring legends. **b)** Taxonomic distribution of genomes included in the search (top) and piggyBac containing taxons (bottom). **c)** Size distribution of N-term, DDE and ZnF containing CRD domain.



Supplementary Figure 3 | Structural and phylogenetic alignment of discovered piggyBac groups. **a)** Representative structures from the 5 mayor PB groups, where the DDE is the main structure and CRD is the additional representation. All of these representatives have been shown to poses activity. All the groups share a similar DDE structure while in CRD structure group 3 has a different fold. **b)** Structural alignment between representative integration domains from the two main ZnF groups, HC6H and C5HC2. The lack of two beta sheets is clearly seen in the C5HC2 PB. **c)** Structural alignment of the two mayor CRD groups together with their respective variants.

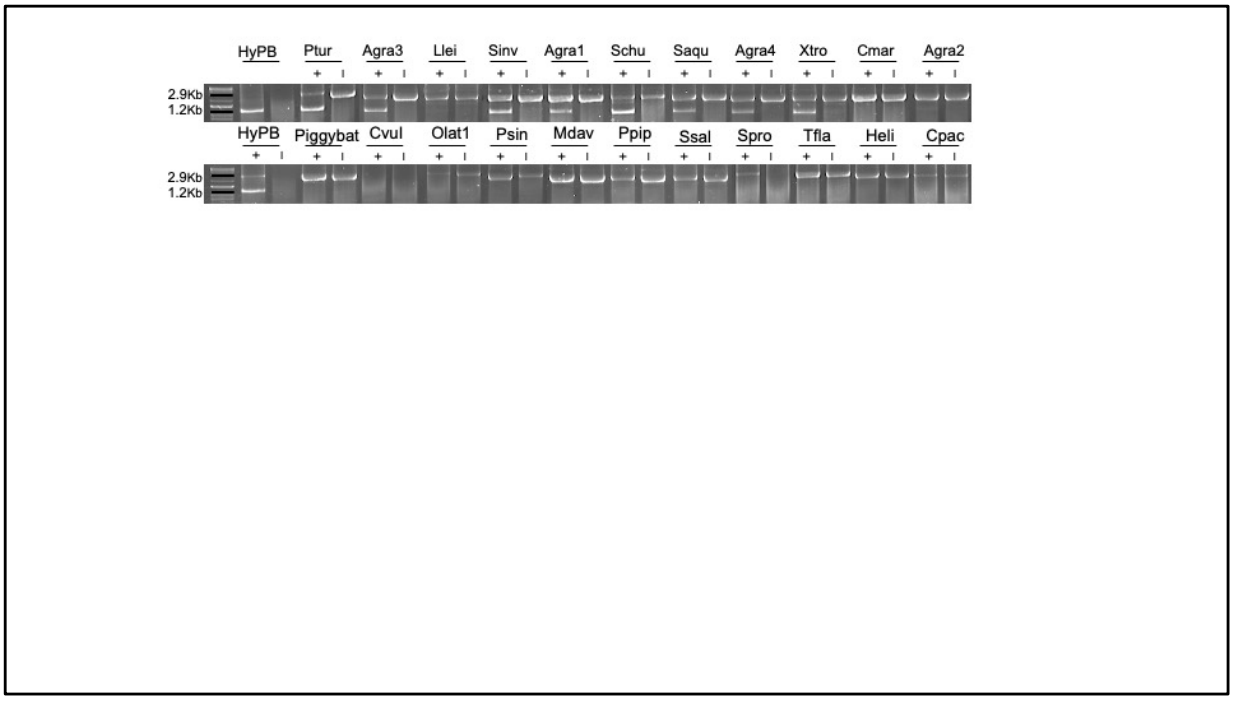


Supplementary Figure 4 | piggyBac superhosts. **a)** Species with most abundant representation of HyPB clusters, termed superhosts. **b)** PiggyBac clusters present in top 3 superhosts (*Homalodisca vitripennis, Prostheceraeus crozeri, Melolontha melolontha*) and the representation across the PiggyBac phylogenetic tree.

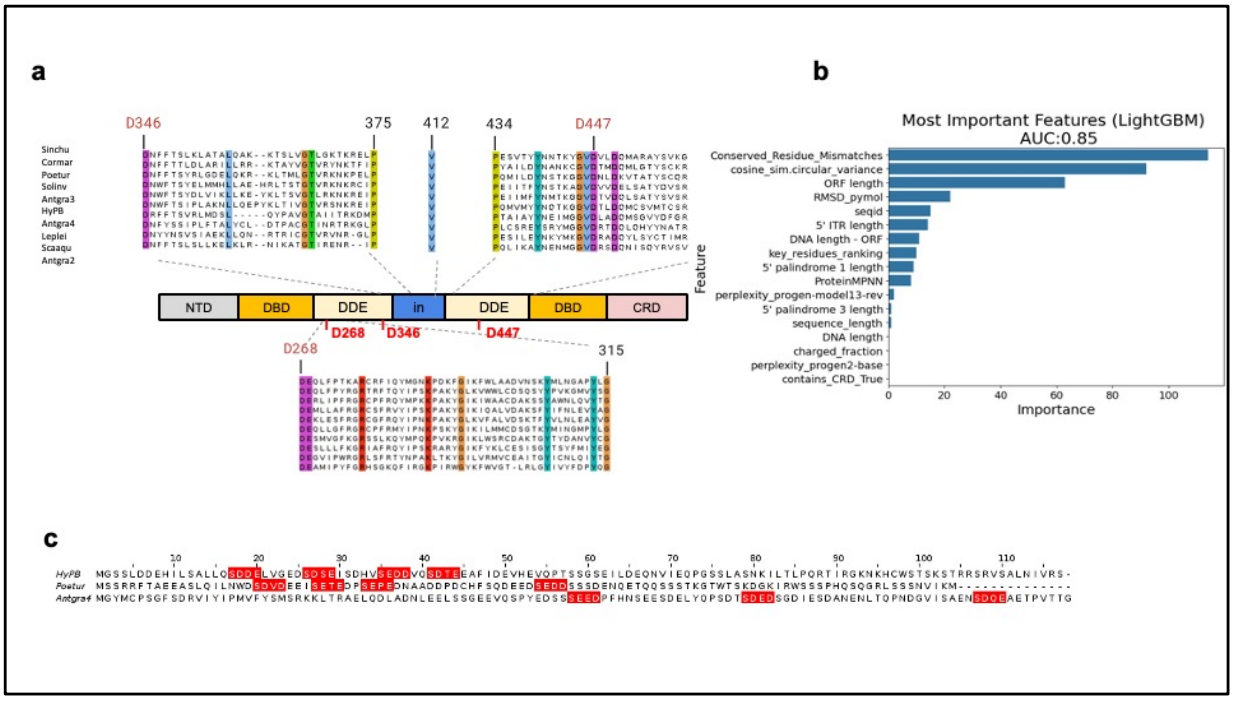


Organized_domains	count	average_length
DNA polymerase phi,DDE	105	587
DDE, DNA replication regu...	95	635.5
DDE, WD40 associated re...	77	595
Zinc finger,DDE	76	602
Herpes virus major out...	71	644
DDE, AF-4 proto-oncoprotein	61	864
Herpes virus major out..., Hamartin protein,DDE,Zin...	49	656
Herpes virus major out...,DDE,Zinc finger	45	648
Atrophin-1 family,DDE	44	848
DDE, Rhodanase C-term...,Zinc finger	36	564
DNA polymerase phi,DDE,Zinc finger	33	575
Anaphase-promoting co...,DDE	32	615
Hamartin protein,DDE,Zinc finger	31	648
Pneumovirinae attac...,Mediator complex subunit...,D...	29	651
YL1 nuclear protein,DDE	23	580
Mitochondrial impor...,DNA polymerase phi,DDE	22	584
Merozoite surface p...,DDE	19	602
Spumavirus gag pro...,Gametogenetin,DDE,Zinc fin...	19	609
DDE, Proline-rich	17	541
YL1 nuclear protein...,Nop14-like family,DDE	17	586
Not1 N-terminal domain,DDE	13	626
Mitochondrial impor...,Zinc finger	12	614.5
DDE, Acetyl-CoA carboxylase	12	573
Procyclic acidic rep...,Surface protein,Ribosome...,D...	12	586
Treacher Collins sy...,Metaviral_G glycoprotein,D...	12	675

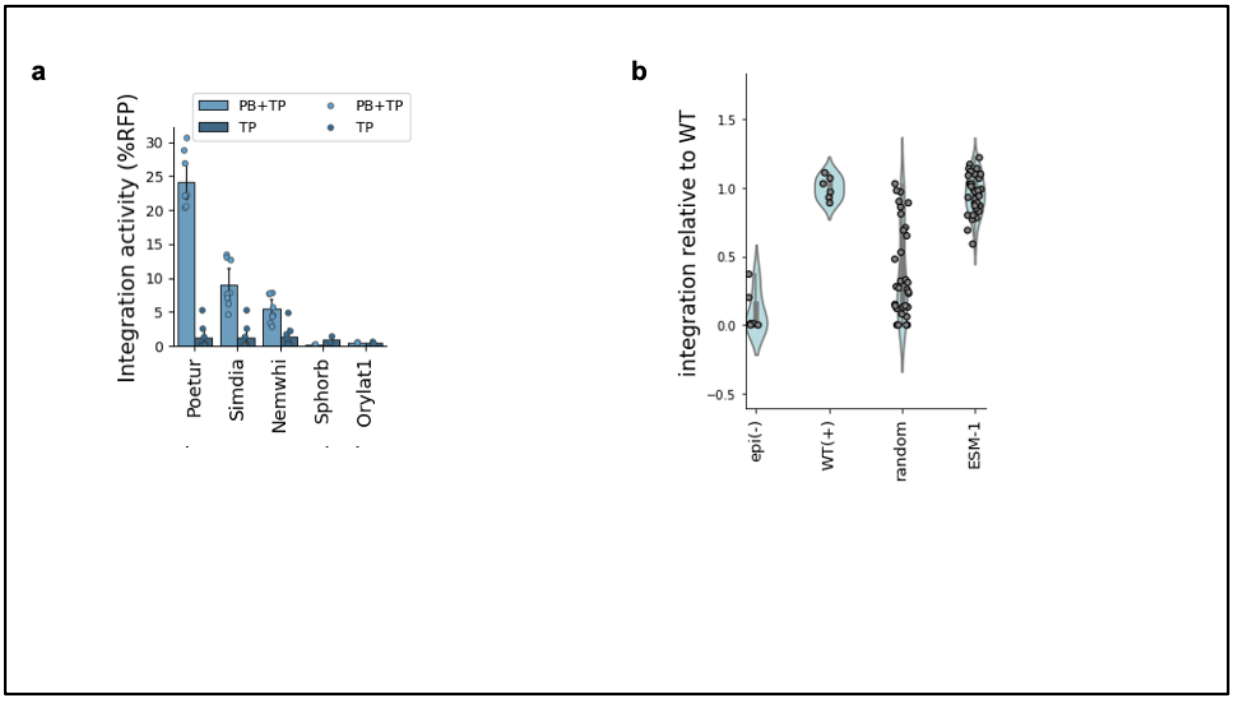
Supplementary Figure 5 | Architectures of piggyBac transposases with fused domains. Transposase fused domain architecture and abundance.



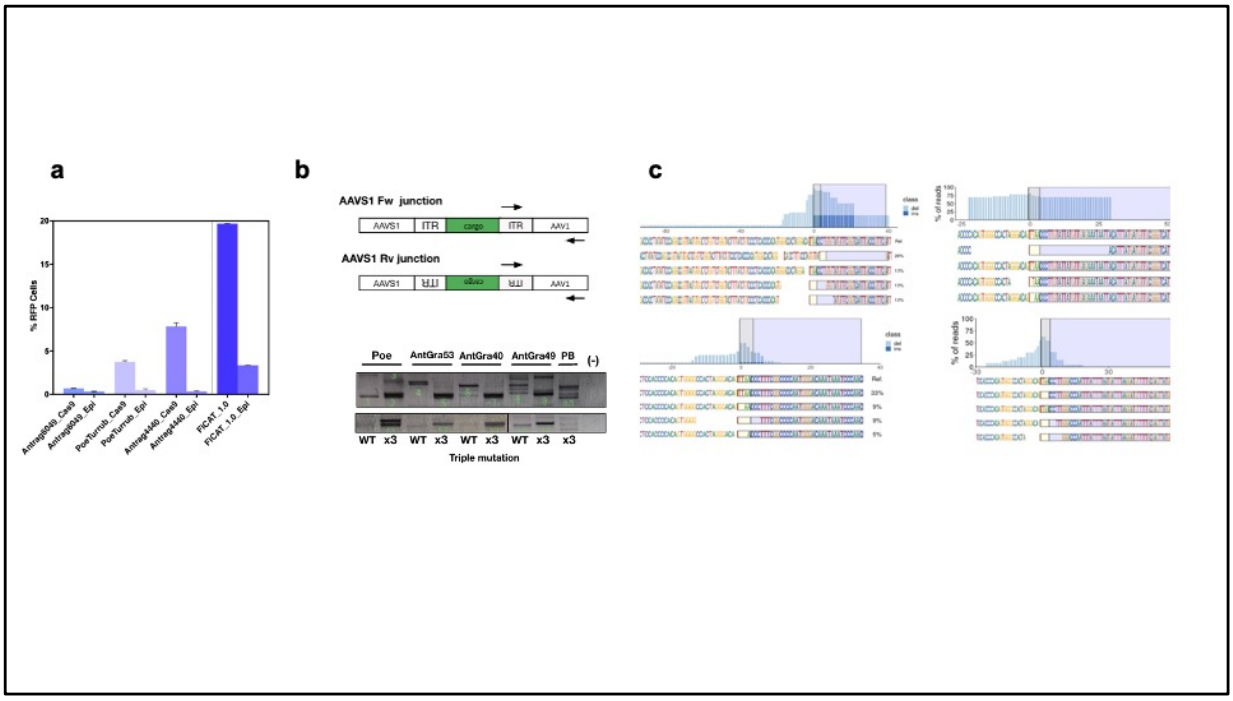
Supplementary Figure 6 | DNA integration activities of active and inactive tested transposons. Excision activity in the 25 transposons selected for experimental testing. Excision is measured by amplification of the transposon donor plasmid after transfection and isolation in HEK293T cells in presence (+) or absence (-) of corresponding transposase. 2.9Kb and 1.2kb bands, correspond to excised and non excised transposon fragment, respectively.



Supplementary Figure 7 | Alignment of active piggyBacs and activity predicting model a) Sequence alignment of active transposase sequences shows key relevant residues. Alignment is truncated to show conserved regions, full alignment in supplementary figure 2. **b)** Activity prediction LightGBM model built with aminoacid, DNA, and computational features to identify features contributing to activity of active vs inactive models. **c)** N-terminal domain showing CKII phosphorylation sites in HyPB, Poetur and Antgra4.



Supplementary Figure 8 | BLAST search and ESM1v guided mutagenesis of Poetur a) Integration levels of Poetur homolog sequences identified by Blast search and filtered. . **b)** Relative long term transposition activity of randomly or ESM1v selected mutants, compared to WT (+) or absence of transposase, epi (-) . 12 ESM1v predicted top scoring mutants in a “zeroshot model” were selected.

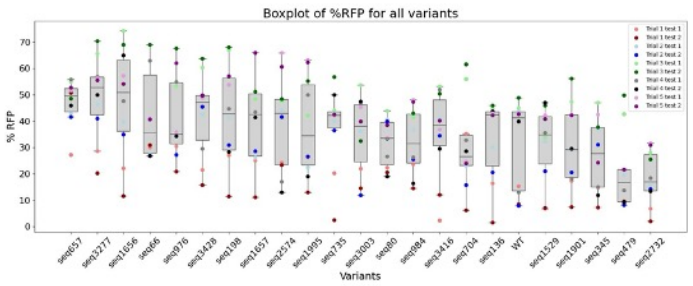


Supplementary Figure 9 | Deployment of discovered transposases as FiCAT precise gene writing tools. Top 3 PiggyBac Transposases identified in the bioprospecting screen were evaluated for programmable integration by co-transfection of Cas9, AAVS1 targeting gRNA and transposase-transposon plasmid pairs. Both WT sequence of piggybac ortholog and HyPB R372A/K375A /D450N excsition+ integration- triple mutant equivalents were tested. (R372A/K375A /D450N are the most important residues for increasing on-target in FiCAT in respect to the HyPB sequence). **a)** RFP levels 3 weeks after transfection in HEK293T cells, representing stable integration. **b)** Junction PCR at the AAVS1-3 target site. **c)** NGS readout at the target site.

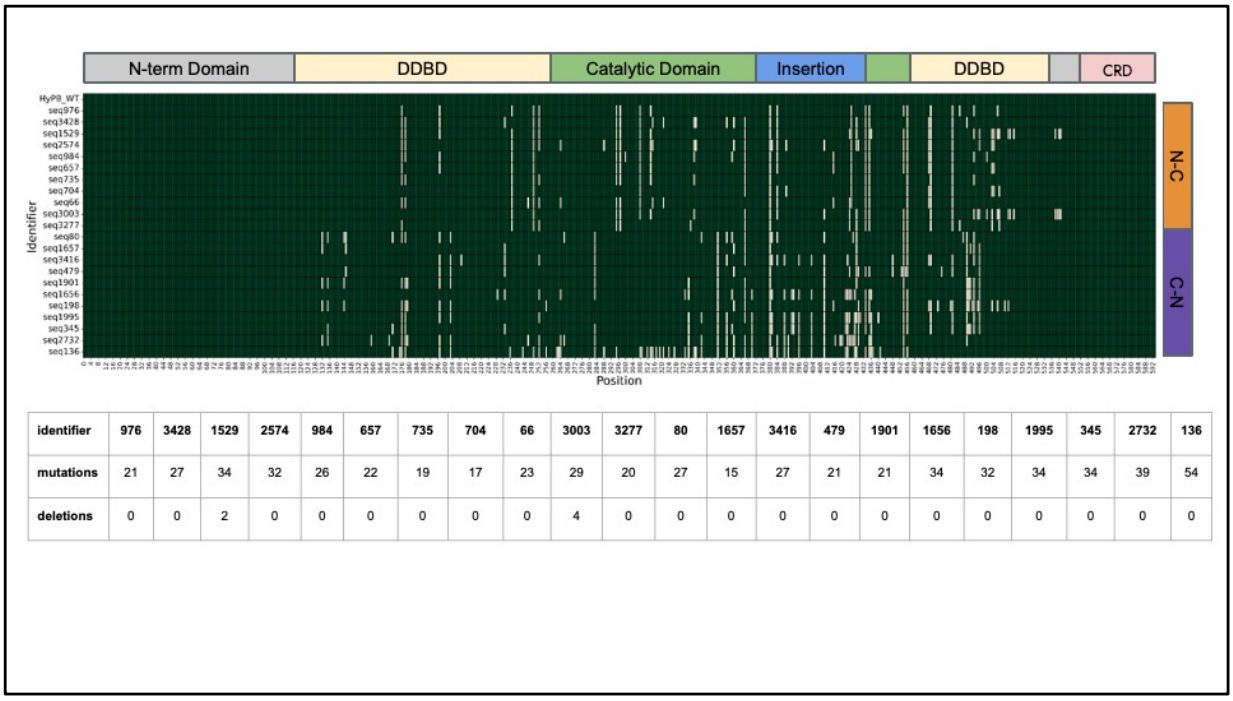
a

Property	N-C Model	C-N Model
Starting	50281	50281
After Removing Duplicates	50281	50237
After Non Canonical Filter	50269	49550
After Repeated Motifs Filter	47530	48517
After CRD Filter	45137	48517
After SeqID Filter	43514	28129
After Catalytic Site Filter	43513	27913
After Alpha bridge filter	43510	27913
After Hyperactive Residue filter	1140	2019
After Key Residue Filter	1115	2018

b

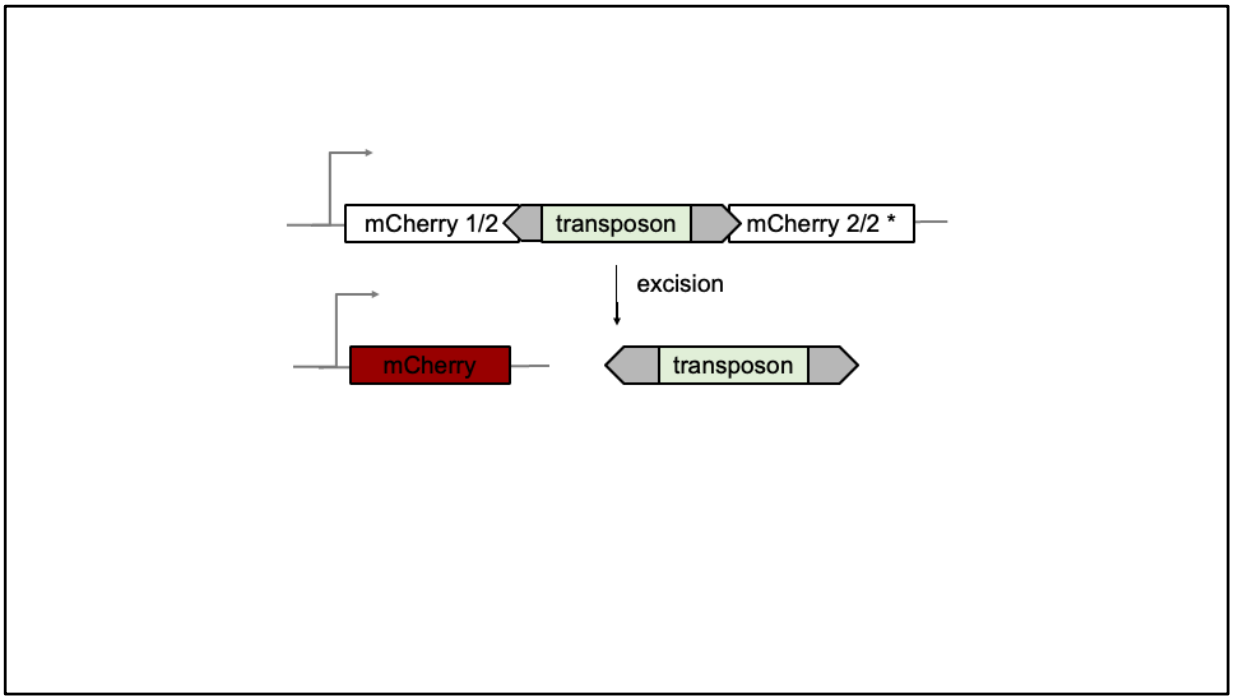


Supplementary Figure 10 | AI generated variants. a) Table of the criteria used to filter out low quality sequences from the set of progen generated sequences. In total 3133 sequences were remaining between the two models after these initial filtering steps. **b)** Shows all experimental %RFP values for the progen-generated sequences that were experimentally evaluated. Sequences are plotted on the y-axis and %RFP values are expressed on the Y-axis. The boxplot shows the 25% and 75% quartiles along with the median value.The points represent the different replicates and are color coded by replicate and trial.



identifier	976	3428	1529	2574	984	657	735	704	66	3003	3277	80	1657	3416	479	1901	1656	198	1995	345	2732	136
mutations	21	27	34	32	26	22	19	17	23	29	20	27	15	27	21	21	34	32	34	34	39	54
deletions	0	0	2	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0

Supplementary Figure 11 | Divergence of LLM generated sequences. Heatmap displaying the positions of all mutations included in the progen-generated variants. Dark green represents unchanged positions and white represents the positions that have a mutated amino acid or a deletion. The Legend above is colored by structurally disordered regions (grey) dimerization and DNA-binding domains (yellow), catalytic domains (green), insertion domain (blue) and the cysteine rich domain (pink). The N-terminal Domain and CRD are unchanged because they were added back to the progen-generated variants. The sequence are ordered by which model they pertain to N-C based (orange) or C-N based (purple) model.



Supplementary Figure 12 | Fluorescence excision assay. Scheme of the fluorescence excision assay used to estimate transposase activity in HEK293T cells.