nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-025-02816-4

Discovery and protein language model-guided design of hyperactive transposases

Received: 27 November 2024

Accepted: 19 August 2025

Published online: 02 October 2025

Check for updates

Dimitrije Ivančić[1,2,6], Alejandro Agudelo[1,2,6], Jonathan Lindstrom-Vautrin[1], Jessica Jaraba-Wallace[1], Maria Gallo[1], Ravi Das[1], Alejandro Ragel[1], Jorge Herrero-Vicente[1], Irene Higueras[2], Federico Billeci[1], Marta Sanvicente-García[1], Paolo Petazzi[1], Noelia Ferruz[3,5], Avencia Sánchez-Mejías[1] & Marc Güell[1,2,4]

The diversity and biochemical potential of the *PiggyBac* transposase gene insertion system remains largely unexplored. Using a eukaryotic transposon mining pipeline, we expand the explored diversity by two orders of magnitude and experimentally validate a subset of highly divergent *PiggyBac* sequences. Fine-tuning a protein language model to further expand *PiggyBac* sequence space discovers transposases with improved activity and that are compatible with T cell engineering and Cas9-directed transposase-assisted integration.

The advancement of genome-engineering technologies has transformed biological engineering and opened new avenues for therapeutic and biotechnological applications[1]. Central to these developments are tools that enable efficient insertion of large DNA sequences into target genomes, an essential capability to unlock the full potential of synthetic biology[2,3]. Among these tools, DNA transposons have been widely adapted for genome modification across numerous organisms[4,5]. Notably, the *PiggyBac* transposase has emerged as a powerful tool because of its ability to integrate substantial DNA cargo across diverse cellular environments, making it a highly versatile platform for gene insertion.

Active *PiggyBac* elements have been identified in the genomes of insects and bats[6,7] and phylogenetic studies have identified *PiggyBac* transposases across multiple eukaryotic families[8,9]. Nonetheless, much of their evolutionary diversity and biochemical potential remain unexplored. Traditionally, exploring *PiggyBac* diversity can be achieved by bioprospecting natural sequences. However, recent advances in generative artificial intelligence (AI) methods applied to protein design have shown that sampled natural diversity can be augmented to generate functional sequences not seen in nature[10–12]. For instance, a combination of RFdiffusion[13] and methodologies to design catalytic sites created active synthetic serine hydrolases with new folds[14]. A protein large language model (pLLM) was recently used to generate a CRISPR–Cas9 that does not exist in nature but performs well for gene-editing applications[10]. The development of such models has opened up exciting opportunities to expand biodiversity and improve gene integration tools. Despite this broad exploration, the potential of *PiggyBac* as a gene insertion tool remains constrained by its preference for TTAA integration sites, limiting its target specificity and precision[15]. Efforts to improve targeting precision have explored fusions with engineered DNA-binding domains such as transcription activator-like effector, engineered zinc-finger proteins and CRISPR catalytically inactive Cas9, each with varying targeting efficiencies[16–18]. Our phylogenetic mining uncovered over 13,000 *PiggyBac* elements, revealing domain acquisitions across multiple *PiggyBac* clusters. We experimentally validated a subset of these elements, identifying ten active transposases with up to 30% sequence identity to one another, thereby expanding the functional repertoire of known *PiggyBac* elements. Additionally, we generated 'mega-active' synthetic variants of the widely used laboratory-evolved hyperactive *PiggyBac* (HyPB) transposase using a fine-tuned pLLM, Progen2 (ref. 19), and demonstrated the applicability of these *PiggyBac* orthologs in critical gene-editing contexts, such as primary T cell engineering and Cas9-directed transposase-assisted integration.

[1]Integra Therapeutics, Barcelona, Spain. [2]Department of Medicine and Life Sciences, Universitat Pompeu Fabra, Barcelona, Spain. [3]Center for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. [4]ICREA, Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain. [5]Present address: Universitat Pompeu Fabra, Barcelona, Spain. [6]These authors contributed equally: Dimitrije Ivančić, Alejandro Agudelo. e-mail: dimitrie.ivancic@upf.edu; marc.guell@upf.edu

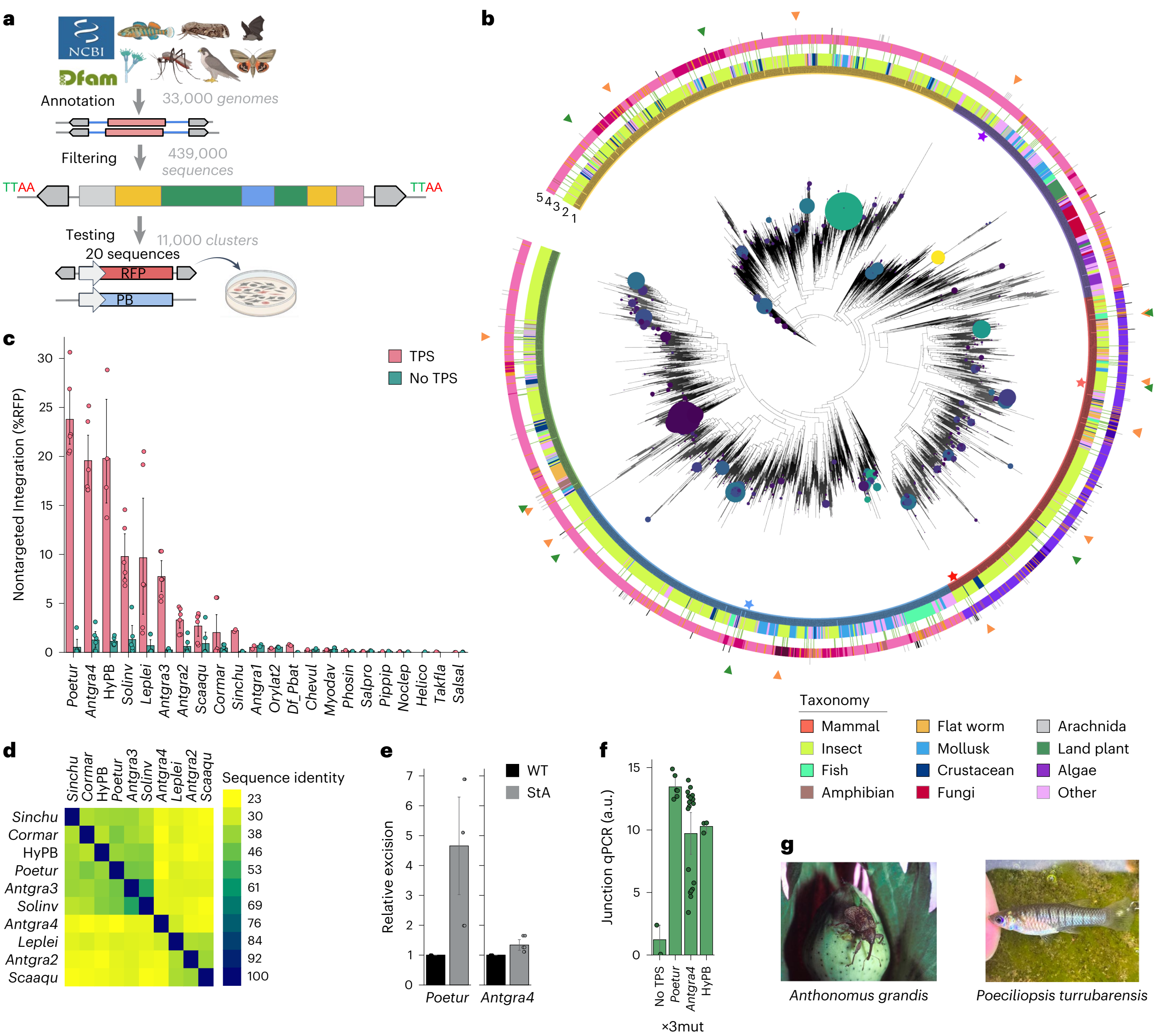

Fig. 1 | *PiggyBac* bioprospecting. a, *PiggyBac* identification and testing pipeline overview (detailed pipeline in Supplementary Fig. 1 and Methods). *Piggybac* domains: N terminus, gray; double DNA-binding domain, yellow; catalytic domain, green; insertion domain, blue; CRD, pink (detailed domain depiction in Supplementary Fig. 7). Panel **a** created with BioRender. **b**, PiggyBac phylogenetic tree from the 2,500 identified clusters at 0.6 identity. Cluster size is represented by the circle radius on top of tree leaves and the number of unique taxonomic species present in the cluster is shown by circle color. Tree ring labels, from inner to outer: (1) identified *PiggyBac* main groups (five in total); (2) major cluster taxonomic groups; (3) clusters with more than one broad taxonomic group; (4) CRD classification; and (5) clusters with fusion domains. Tested *PiggyBac* clusters are marked with arrows, inactive *PiggyBac* clusters are marked with orange arrows and active *PiggyBac* clusters are marked with green arrows. The four colored stars represent previously described *PiggyBac*-like transposons with demonstrated autonomous activity: *PiggyBat*[6], blue; *PiggyBac*[5], red; *Mage*[32], orange; *PLE-wu*[33], purple. The 'fish' category includes *Chondrichthyes*, *Agnatha* and *Osteichthyes* (complete legends and colors in Supplementary Fig. 2). **c**, Experimental validation of *PiggyBac* orthologs by nontargeted transposon integration fluorescence assay in HEK293T cells 2 weeks after transfection, in the presence (TPS, pink) or absence (no TPS, green) of transposase plasmid. Data are presented as the mean values ± 95% confidence interval (CI), with $n = 2$ for orthologs with a mean level of RFP lower than 1% and $n = 3$ for those with higher (seven top performers). **d**, Sequence identity heat map between active orthologs from **c**. **e**, Effect of N-terminal phosphorylation substitutions on excision, measured by transposon excision fluorescence assay. StA indicates serine-to-alanine substitutions in CKII phosphorylation sites (Supplementary Fig. 7). Data are presented as the mean values relative to WT ± 95% CI, with $n = 3$. **f**, Targeted transposon integration qPCR assay with *Poetur* and *Antgra4* orthologs in the triple-mutant background (R372A;K575A;D450N) at the AAVS1-3 site. Data are presented as the mean values ± 95% CI, with $n = 1$. **g**, Pictures of species containing the top two *PiggyBac* hits[34]. Credits: *A. grandis*, photo courtesy of USDA Agricultural Research Service; *P. turrubarensis*, Paradise Costa Rica.

We searched all available eukaryotic genome assemblies on the National Center for Biotechnology Information (NCBI; 31,565 genomes) and Dfam[20] (20,638 *PiggyBac* sequences) databases, finding a total of 273,643 *PiggyBac* transposon open reading frames (ORFs) together with their DNA sequences (Fig. 1a and Supplementary Fig. 1). To differentiate active transposons from transposase-derived proteins co-opted by the host that have lost transposition activity[21,22], we retrieved sequences with the presence of an RNase H-like domain, cysteine-rich domain (CRD), terminal inverted repeats (TIR) and a target site duplication (TSD) with the TTAA motif (Supplementary Fig. 1). These motifs are reported to be crucial for DNA excision and integration[7]. Filtering yielded a dataset of 116,216 putatively transposition competent *PiggyBac* elements that resulted in 13,693 PiggyBac subfamilies after clustering at 80% sequence identity.

The eukaryotic distribution of *PiggyBac* transposons is notably diverse, encompassing taxa from fungi and plants to mammals (Fig. 1b and Supplementary Fig. 2b); it is predominantly represented in insects (~60%), followed by fish and mollusks (5%). We identified five main *PiggyBac* groups (Fig. 1b and Supplementary Figs. 2a and 3a) on the basis of main tree phylogenetic branches, taxonomic distribution and the CRD types. More than 200 clusters are represented by more than one broad taxonomic group (Fig. 1b, ring 3), indicating widespread horizontal gene transfer across groups, as previously reported in other transposable elements[23]. Group 4 has a unique, unexpected taxonomic distribution with presence in fungi, land plants and algae (Fig. 1b, ring 1, purple). We also observed 'superhost' species, characterized by containing numerous *PiggyBac* sequences. The top three superhosts captured 7.3% of all *PiggyBac* diversity (Supplementary Fig. 3). Additionally, we found multiple domain acquisition events at both N and C termini, with 4.6% of all the reported clusters containing a fusion domain and N-terminal fusions being more predominant (Fig. 1b, ring 5). DNA-binding domains and fusogens were the most abundantly acquired domains, suggesting multiple transposition mechanisms for DNA recognition and cell entry (Supplementary Fig. 4).

We used AlphaFold3 (ref. 24) structural prediction and clustering to further understand the diversity of the CRD domain. We identified two main CRD cross brace zinc-finger folds, HC6H and C5HC2 (Supplementary Fig. 5). In contrast to C5HC2, the HC6H group is longer and retains two unique β-sheets in its insertion domain. The insertion domain consists of structures with three and five β-strands in C5HC2 and HC6H, respectively, which interrupts the catalytic domain after the seventh β-strand. While the catalytic domain catalyzes the hydrolysis and transesterification steps necessary for transposition, the insertion domain has a role in DNA binding and transposon integration[7]. Analysis of the catalytic domain indicates high structural conservation (root-mean-square deviation (r.m.s.d.) of the catalytic region near 2 Å and a template modeling (TM)-score of 0.915) despite high sequence divergence (Supplementary Fig. 5).

To explore the potential of bioprospected transposon diversity for gene insertion, we selected 23 representative *PiggyBac* sequences across the phylogenetic tree for experimental testing (Fig. 1b, colored triangles). These sequences were chosen to encompass all five major *PiggyBac* groups, both primary CRD types and a representative range of taxonomic groups. Transposition activity was validated through detecting excision of the transposase plasmid (Supplementary Fig. 6a) and nontargeted integration of a red fluorescent protein (RFP)-containing transposon payload in HEK293T cells (Fig. 1c). Nontargeted integration refers to the canonical *PiggyBac* transposition mechanism, in which it excises and inserts itself into TTAA motifs throughout the genome[25]. Of the tested sequences, nine (~40%) had detectable activity, with two sequences equivalent to laboratory-evolved HyPB[5]. Active sequences were spread across phylogeny and had low sequence identity to HyPB (Fig. 1d). This broad distribution of active elements across taxonomic and CRD diversity underscores the potential of *PiggyBac* transposons as versatile tools in genetic engineering and gene-transfer applications. Interestingly, the previously described PiggyBat sequence did not exhibit activity, which contrasts with previous reports[6]. This discrepancy is likely because of the fact that a consensus *PiggyBat* sequence generated in this study is constructed from multiple *PiggyBat* cluster sequences and is different from the previously described. To further improve transposon activity, we identified and removed CKII phosphorylation motifs in the N terminus of *PiggyBac*, previously reported to inhibit its transposition activity in HyPB[7] (Supplementary Fig. 7b). CKII site removal increased transposition activity in both orthologs (Fig. 1e). We also tested how TIR truncation affected excision in *Poetur* and *Antgra4* (Supplementary Fig. 8), identifying minimal TIR versions with equal activity. We further tested compatibility of our orthologs with the previously described FiCAT[18] targeted insertion system. In the FiCAT platform, a Cas9 enzyme fused to an engineered *Piggybac* transposase induces a double-strand break (DSB) at a target genomic site. The *PiggyBac* component, engineered to be excision competent and integration deficient, excises a transposon delivered by plasmid. This transposon is then inserted into the DSB site, generating an integration signature mediated by nonhomologous end joining. Our results showed successful FiCAT compatibility of *Poetur* and *Antgra4* in HEK293T cells (Fig. 1f and Supplementary Figs. 9 and 10).

Next, we sought to explore how the generated corpus of natural sequences could be used to improve the activity of existing transposases. We fine-tuned the ProGen2-base language model[19] using over 13,000 bioprospected sequences, similarly to the method previously described for Cas9 nucleases[10]. In our training data, the HyPB sequence was included five of ten times, depending on the model, to bias the model toward improvement of the HyPB sequence. We created two separate models: one model to generate sequences from the N terminus to C terminus and the second to generate sequences from the C terminus to N terminus. We then generated over 100,000 sequences from these two models prompted with the first 50 (N–>C) or last 50 (C–>N) amino acids. A total of 50 amino acids were selected to give sufficient context to the models so that they could generate similar sequences, without giving so much that the model could perfectly recreate the HyPB sequence. Sequences were first filtered on the basis of a set of basic protein properties in addition to *PiggyBac*-specific properties (Fig. 2a and Supplementary Fig. 8b). We further filtered and scored sequences by structural (predicted local distance difference test (pLDDT), r.m.s.d. to experimental structure, SURFMAP[26,27] and TM-scores) and deep learning scores (Progen perplexity, ProteinMPNN[28] and ESM1v[29]). Generated sequences had higher pLDDT, ESM1v and ProteinMPNN scores when compared to a matched subset of natural sequences, indicating that the designed sequences may have higher activity than the natural ones (Fig. 2b). ESM1v is a pLLM developed by Meta Research that was designed for predicting variant effects, ProteinMPNN is a deep learning-based sequence design method that can decode amino acid sequences from structural representations of proteins and score proteins and pLDDT is a metric used by structural prediction tools to evaluate the confidence of predictions. These metrics have previously been used for computational scoring of enzymes[13].

We experimentally tested 11 sequences from each model (22 total), 15–54 mutations apart from the original HyPB sequence. All of the generated sequences displayed excision activity with an average percentage RFP ranging from 15% to 48% excision (Supplementary Fig. 11c). Of the tested sequences, seven of 22 were significantly more active in excision than the laboratory-evolved HyPB (Fig. 2c) (Mann–Whitney *U*-test with a *P*-value cutoff of 0.05). We further evaluated nontargeted integration of the synthetic sequences (Supplementary Fig. 11d). seq3277 was the most active sequence in both excision and nontargeted integration. We termed this sequence Mega-*PiggyBac*. Curiously, seq136 showed the highest nontargeted integration efficiencies while having baseline excision activities and had the highest number of substitutions (54 amino acids (aa)), most of them in the catalytic region. To evaluate the relevance of the proposed pLLM-based sequence improvement approach, we tested both bioprospected sequences near the *Poetur* sequence space and single mutants predicted to have improved fitness by ESM1v ('zero-shot' approach[29]) as comparable optimization approaches. In contrast to pLLM, none of these approaches led to mutants with significantly increased nontargeted integration activity (Supplementary Fig. 12).

We gathered multiple metrics to both inform our selection and aid post hoc learning of properties associated with transposase activity.

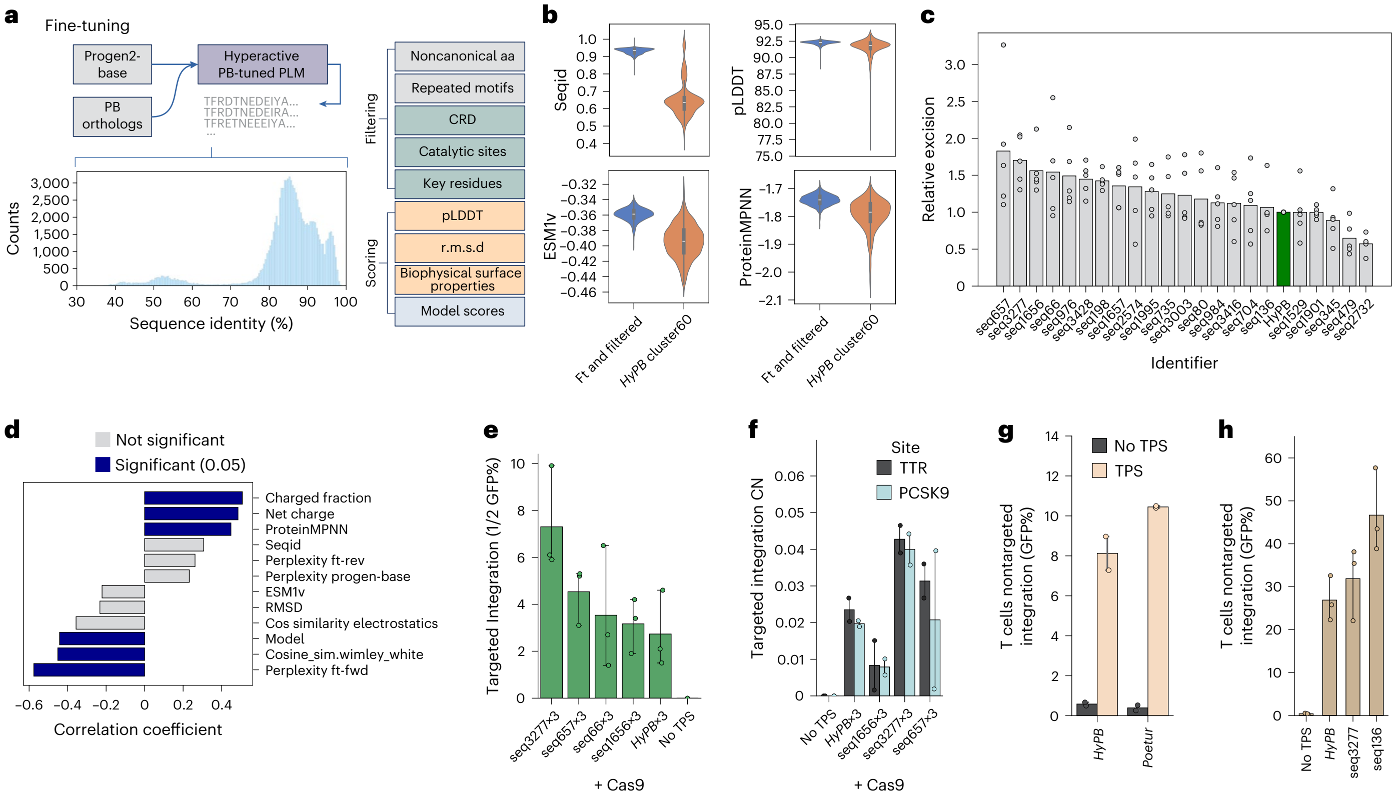

Fig. 2 | Synthetic mega-active *PiggyBac* generation using protein language model fine-tuning. a, Overview of the fine-tuning and sequence generation pipeline. The Progen2-base model was fine-tuned on a set of over 10,000 *PiggyBac* orthologs identified through the bioprospecting pipeline. Over 100,000 sequences were generated with a sequence identity between 35% and 99% to the HyPB. Sequences were then filtered using a set of basic (gray) and *PiggyBac*-specific (green) amino acid sequence metrics and scored using a set of scores based on structural (orange) and deep learning (blue) metrics to select a final subset of 22 sequences for experimental validation. **b**, Distribution of four key metrics (sequence identity, pLDDT, ProteinMPNN score and ESM1v score) for natural sequences from the *HyPB* cluster at 60% identity (orange) and sequences generated from our progen-ft model (blue) after filtering. The violin plots represent the entire distribution of scores for the two sets of sequences and the internal box plot represents the quartiles for each score, with the center being the median, the bottom and top being the first and third quartiles, respectively, and the whiskers going 1.5× the interquartile range from the top and bottom. Ft, Fourier transform. **c**, Relative excision for progen-ft-generated variants normalized to HyPB activity (highlighted in green), measured by a transposon excision fluorescence assay. Bars reflect the mean relative excision over the four trials and points represent the mean relative excision of replicates in each trial. Data are presented as the mean values, with $n = 5$. **d**, Correlations between calculated and measured features to relative excision of the progen-ft-generated variants. Significant correlations are highlighted in dark blue. Correlation was measured with Pearson's correlation. **e**, Targeted integration with top pLLM-generated mutants, measured by a targeted transposon integration GFP reconstitution assay that measures integration of a 1/2 GFP reporter cargo upstream of a stably integrated 2/2 GFP in HEK293T reporter cell line. Triple-mutant (×3) versions of the transposases were made by selecting the residues corresponding to R372A;K375A;D450N in HyPB. Data are presented as the mean values ± 95% CI, with $n = 3$. **f**, Targeted transposon integration measured by digital PCR assay in C2C12 mouse myoblast cell lines at *TTR* and *PCSK9* loci for top AI-designed transposases. The sum of integration in both orientations is shown. Data are presented as the mean values ± 95% CI, with $n = 2$. **g**, Nontargeted transposon integration measured by fluorescence assay in primary T cells for top bioprospected ortholog *Poetur* 7 days after electroporation. Data are presented as the mean values ± 95% CI, with $n = 2$. **h**, Nontargeted integration of a GFP cargo in primary T cells with *HyPB* and top synthetic sequences transposases 7 days after cell electroporation. Data are presented as the mean values ± 95% CI, with $n = 3$.

The structural and AI-based scores described above were used to help guide our final selection and, following experimental testing of our variants, certain metrics were found to be correlated to transposase activity. Net charge of the protein, charged fraction of amino acids (ratio of charged amino acids in the sequence) and ProteinMPNN score seemed to be positively correlated with protein activity. In contrast, perplexity scores from the N–>C fine-tuned model, model version (N–>C or C–>N) and Wimley–White[30] surface structural similarity scores seemed to be negatively correlated (Fig. 2d and Supplementary Fig. 10a).

We then tested top hits for FiCAT targeted integration (Fig. 2e). We found that synthetic sequence 3277 improved targeted integration twofold, demonstrating that improved pLLM-generated sequences are compatible with programmable gene insertion. We further validated targeted integration with top pLLM-generated sequences in mouse c2c12 myoblast cells at *TTR* and *PCSK9* loci (Fig. 2f). To illustrate the potential impact of bioprospecting guided sequence discovery for therapeutic applications, we stably delivered a GFP transposon cargo with *Poetur* and AI-designed transposases in T cells, showing higher nontargeted integration for *Poetur* (Fig. 2g) and for seq136 (Fig. 2h) when compared to *HyPB*, while seq3277 (Fig. 2h) had same nontargeted integration activity despite having higher excision and targeted integration, underscoring that diversity in pLLM-generated sequences can capture optimization toward different protein properties.

Our work expands the phylogenetic tree of *PiggyBac* transposons by two orders of magnitude, unveiling a previously unexplored diversity within this family of mobile genetic elements. This expansion led to the discovery and characterization of nine additional active *PiggyBac* orthologs, broadening the range of transposase variants available for research and biotechnological applications. Among these identified orthologs, two stand out for their exceptional performance, demonstrating activity levels comparable to those of evolved HyPB variants and robust activity in primary T cells, an essential target for many therapeutic applications in gene and cell therapy. Importantly, the discovered orthologs are compatible with the FiCAT programmable gene insertion system. This compatibility paves the way for innovative approaches to gene insertion, enhancing the system's versatility in applications ranging from gene therapy to synthetic biology. Furthermore, we exemplified how pLLM de novo sequence generation offers a powerful approach to improving transposase activities. This method enhances the optimization process and provides a framework where the modifications are informed by a comprehensive sequence–function relationship. By leveraging the capabilities of pLLM, researchers could use the described method to systematically identify variants with enhanced properties.

Recent work demonstrated substantial activity improvement upon TIR truncation[31]. Moreover, combining this knowledge on TIR architecture with recently developed genome language models could further improve transposition activity. Additionally, determining how AI-guided activity improvement impacts specificity will be crucial for successfully using these methods for therapeutic protein development.

Our findings underscore the power of combining bioprospection with AI-driven sequence optimization to accelerate the discovery and enhancement of next-generation gene insertion tools. This approach not only expands the *PiggyBac* toolkit but also provides a valuable framework for the development of additional gene modification tools for precise and efficient genome manipulation applicable across biotechnology and therapeutic fields.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-025-02816-4.

References

1. Wang, J. Y. & Doudna, J. A. CRISPR technology: a decade of genome editing is only the beginning. *Science* **379**, eadd8643 (2023).
2. Yarnall, M. T. N. et al. Drag-and-drop genome insertion of large sequences without double-strand DNA cleavage using CRISPR-directed integrases. *Nat. Biotechnol.* **41**, 500–512 (2023).
3. Mukhametzyanova, L. et al. Activation of recombinases at specific DNA loci by zinc-finger domain insertions. *Nat. Biotechnol.* **42**, 1844–1854 (2024).
4. Li, X. et al. *PiggyBac* transposase tools for genome engineering. *Proc. Natl Acad. Sci.* **110**, E2279–E2287 (2013).
5. Yusa, K., Zhou, L., Li, M. A., Bradley, A. & Craig, N. L. A hyperactive *PiggyBac* transposase for mammalian applications. *Proc. Natl Acad. Sci. USA* **108**, 1531–1536 (2011).
6. Mitra, R. et al. Functional characterization of *PiggyBat* from the bat *Myotis lucifugus* unveils an active mammalian DNA transposon. *Proc. Natl Acad. Sci. USA* **110**, 234–239 (2013).
7. Chen, Q. et al. Structural basis of seamless excision and specific targeting by *PiggyBac* transposase. *Nat. Commun.* **11**, 3446 (2020).
8. Yuan, Y.-W. & Wessler, S. R. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl Acad. Sci. USA* **108**, 7884–7889 (2011).
9. Guo, M. et al. *PiggyBac* transposon mining in the small genomes of animals. *Biology* **13**, 24 (2024).
10. Ruffolo, J. A. et al. Design of highly functional genome editors by modeling the universe of CRISPR–Cas sequences. *Nature* https://doi.org/10.1038/s41586-025-09298-z (2025).
11. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
12. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322 (2019).
13. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
14. Lauko, A. et al. Computational design of serine hydrolases. *Science* **388**, eadu2454 (2025).
15. Galvan, D. L. et al. Genome-wide mapping of *PiggyBac* transposon integrations in primary human T cells. *J. Immunother.* **32**, 837–844 (2009).
16. Luo, W. et al. Comparative analysis of chimeric ZFP-, TALE- and Cas9-*PiggyBac* transposases for integration into a single locus in human cells. *Nucleic Acids Res.* **45**, 8411–8422 (2017).
17. Hew, B. E., Sato, R., Mauro, D., Stoytchev, I. & Owens, J. B. RNA-guided *PiggyBac* transposition in human cells. *Synth. Biol.* **4**, ysz018 (2019).
18. Adrian, K. et al. RNA-guided retargeting of *Sleeping Beauty* transposition in human cells. *eLife* **9**, e53868 (2020).
19. Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N. & Madani, A. Progen2: exploring the boundaries of protein language models. *Cell Syst.* **14**, 968–978 (2023).
20. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* **12**, 2 (2021).
21. Bouallègue, M., Rouault, J.-D., Hua-Van, A., Makni, M. & Capy, P. Molecular evolution of *PiggyBac* superfamily: from selfishness to domestication. *Genome Biol. Evol.* **9**, 323–339 (2017).
22. Cosby, R. L. et al. Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* **371**, eabc6405 (2021).
23. Zhang, H.-H., Peccoud, J., Xu, M.-R.-X., Zhang, X.-G. & Gilbert, C. Horizontal transfer and evolution of transposable elements in vertebrates. *Nat. Commun.* **11**, 1362 (2020).
24. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
25. Yusa, K. *PiggyBac* transposon. *Microbiology Spectrum* **3**, MDNA3–0028–2014 (2015).
26. Schweke, H., Mucchielli, M.-H., Chevrollier, N., Gosset, S. & Lopes, A. SURFMAP: a software for mapping in two dimensions protein surface features. *J. Chem. Inf. Model.* **62**, 4211–4219 (2022).
27. Sanner, M. F., Olson, A. J. & Spehner, J.-C. Reduced Surface: an efficient way to compute molecular surfaces. *Biopolymers* **38**, 305–320 (1996).
28. Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
29. Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.09.450648 (2021).
30. Wimley, W. C. & White, S. H. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nat. Struct. Biol.* **3**, 842–848 (1996).
31. Hickman, A. B. et al. Activity of the mammalian DNA transposon *PiggyBat* from *Myotis lucifugus* is restricted by its own transposon ends. *Nat. Commun.* **16**, 458 (2025).
32. Tian, J. et al. Mage transposon: a novel gene delivery system for mammalian cells. *Nucleic Acids Res.* **52**, 2724–2739 (2024).
33. Wu, C. & Wang, S. PLE-wu, a new member of *PiggyBac* transposon family from insect, is active in mammalian cells. *J. Biosci. Bioeng.* **118**, 359–366 (2014).
34. Lindstrom-Vautrin, J. & Agudelo, A. *PiggyBac* bioprospecting pipeline. *GitHub* https://github.com/Integra-tx/Piggybac_bioprospecting_pipeline.git (2025).

Methods

Retrieval of *PiggyBac* transposons

Complete *PiggyBac* transposon sequences were gathered from all available eukaryotic genomes in the NCBI database[35] (31,565 genomes) and all *PiggyBac* elements in the Dfam database (20,638). Dfam sequences were directly downloaded by selecting entries labeled as *PiggyBac*. NCBI eukaryotic genome-derived transposase sequences were identified using Bath[36,37], with a custom hidden Markov model constructed from all active *PiggyBac* sequences reported in the literature. For NCBI PB retrieval, flanking regions 4 kbp upstream and downstream were included to capture the complete transposon sequence including DNA TIRs. A filter was applied to retain *PiggyBac* transposases longer than 250 aa. After this filtering, a total of 273,643 *PiggyBac* were recovered, with a mean transposase length of 500 residues and mean DNA transposon length of 3,298 bp.

To refine the boundaries of each transposon in the NCBI dataset, clustering by RNase H-like domains of the *PiggyBac* hits at a 0.9 similarity threshold was performed with MMseqs2 (ref. 38), followed by multiple-sequence alignment (MSA) of the complete DNA sequences (including flanking regions) within clusters using MAFFT[39]. Transposon boundaries were then delimited on the basis of the MSA results.

Filtering for active *PiggyBac* elements

To identify active *PiggyBac* transposons from all the transposons identified in the previous step, we applied the following sequential filters:

1. RNase H-like domain identification: The presence of a RNase H-like domain was confirmed using RPS-BLAST[40], with the Conserved Domain Database[41] as the reference database and selecting only sequences with an RNase H-like domain longer than 250 aa.
2. CRD identification: A total of 50 representative CRDs were manually curated and structurally modeled using AlphaFold3 (ref. 24) to identify residues directly involved in zinc ion coordination. On the basis of this curated set, we derived a set of sequence motifs (Supplementary Table 2), revealing major CRD groups and their variants. CRDs were then identified using regular expressions matching these curated motifs.
3. TIR identification: TIRs were identified in the flanking DNA regions using the EMBOSS tool Palindrome[42], focusing on pairs of palindromic sequences located on opposite flanks of the transposon in the first and last 200 bp. We retained only TIRs with at least two palindromic sequences of 10 bp or longer and allowing up to two mismatches. As an additional quality control step, only palindromes in which the two most common nucleotides account for less than 80% of the palindrome were selected.
4. TSD identification: TSDs were searched for with regular expression within the first and last 50 bp of each transposon, using the motif TTAACC, with up to two allowed mismatches.

A total of 116,216 putatively active *PiggyBac* elements were recovered after applying the filtering process.

Dataset clustering

The filtered dataset was then clustered to reduce redundancy using the RNase H-like domain of the transposase. We performed two clusterings with MMseqs2, one at 0.8 identity and one at 0.6 identity. The 0.8 clustering was performed following transposon annotation 80–80–80 (ref. 43), as it is considered that two transposon elements belong to the same family if they share 80% (or more) sequence identity in at least 80% of their coding or internal domain. This dataset was used for the fine-tuning of the pLLMs. The clustering at 0.6 was performed to make a broader classification of *PiggyBac* families and used for the phylogenetic analysis. The clustering at 0.8 produced 13,693 clusters, while that at 0.6 produced 2,572 clusters.

Phylogenetic analysis of bioprospected sequences

The phylogenetic tree was built with IQ-TREE (version 1.6.12)[44] on the basis of an MSA generated with the 2,572 centroids from the 0.6 clustering with MUSCLE[45]. Model finder[46] was used to select the optimal model for accurate phylogenetic estimation (LG + R10) and UFBoot[47] was used for bootstrap approximation with 1,000 replicates. The resulting tree was visualized using iTOL[48]. Additional *PiggyBac* domains were identified with RPS-BLAST[40]. Molecular graphics were generated using UCSF Chimera[49].

Blast identification of *Poetur* orthologs

A search with BLASTn on the core nucleotide database was conducted using *Poetur*. The whole transposon, including the TIR and TSD were included to find hits that also possessed these motifs. A total of four hits from four different species were manually selected on the basis of them having a coverage higher than 88%, sequence identity higher than 83% and the presence of all necessary functional domains for transposition activity (RNase H-like domain, CRD, TIR and TSD).

Model fine-tuning

The ProGen2-base[19] language model of 764 million parameters was fine-tuned on over 13,000 sequences from the *PiggyBac* orthologs clustered at 0.8. This fine-tuning was performed to give the ProGen2-base model a better understanding of *PiggyBac* sequences. In this process, the pretrained model was further trained on the *PiggyBac* orthologs and, as the model trained, the 764 million parameters were updated in a way that aimed to minimize the cross-entropy loss. We fine-tuned two separate models: one model to generate sequences from the N terminus to C terminus and the second to generate sequences from the C terminus to N terminus. Both models were fine-tuned using the full amino acid sequences excluding the N-terminal domain, which was excluded because it is an extremely variable domain. In the HyPB, the N terminus consists of the first 116 aa and, in general, the N terminus is a disordered region leading up to the first double DNA-binding domain region.

The sequences were split using a 80:20 train–test split. In addition to the set of orthologous sequences used in the training, additional wild-type (WT) *HyPB* sequences (5–10) were added to the training set to bias the model toward HyPB. This allowed us to generate sequences in a closer sequence identity range to HyPB than we were able to without biasing the dataset. Fine-tuning was performed using the Trainer module fetched from Hugging Face over two epochs with a training batch size of 4 and evaluation batch size of 8. A constant learning rate of 5.0×10^{-5} was used and the model was evaluated after every 2,000 steps. Cross-entropy loss was used to evaluate every checkpoint in the model and the checkpoint with the lowest validation loss was used for sequence generation. The remaining Trainer parameters were kept at the default values. A full exploration of the Trainer hyper parameters was not performed as, with these fairly standard parameters, we were able to generate convincing sequences with our desired properties.

AI sequence generation

In both models, 50 aa from WT HyPB were used to prompt sequence generation. An initial prompt was used to give the model enough context to build a *PiggyBac*-like sequence. In preliminary testing, 50 aa seemed to provide a good balance of giving the models a good starting point without allowing them to replicate the HyPB sequence perfectly. For the N−>C model, the first 50 aa after the N-terminal domain were used and, in the C−>N model, the final 50 aa of the CRD were used to prompt sequence generation. For the C−>N model, sequences were generated 'backward' and then reversed to have the standard directionality. The maximum sequence length for both models was set to 500 aa and a temperature of $T = 0.5$ and nucleus probability $P = 0.95$ were used.

AI sequence filtering

The generated sequences first went through a set of three basic filters. First, duplicated sequences were removed. Second, sequences with noncanonical amino acids were removed. Third, sequences were filtered using a *k*-mer repetition filter such that no amino acid motif of six, four, three or two residues was repeated two, three, six or eight times consecutively. The next set of filters were HyPB specific and included testing for a *PiggyBac* CRD (based on the presence of at least seven cysteine amino acids in the final 50 aa), sequence identity to WT (80–95% to the RNAse H-like and CRD domains) and specific key residues including catalytic site, α-bridge residues, hyperactive residues and another extensive set of key residues including DNA-interacting residues.

For all of these sequences, we calculated perplexity using the ProGen2-base model and the fine-tuned model responsible for generating a given sequence. For a subset of sequences that passed our filters, structures were predicted using ESMFold[50]. Structures were then compared to the experimentally available *PiggyBac* structure (PDB 6X67) to extract r.m.s.d. and TM-scores using PyMOL (Schrödinger) and TMAlign[51], respectively. Finally, structures were aligned to the experimental *PiggyBac* structure and several surface properties were calculated using SURFMAP: a tool that projects surface residues from a protein structure into a two-dimensional space and can calculate different amino acid residue properties. The five metrics we calculated using SURFMAP were stickiness, circular variance, Wimley–White, Kyte–Doolittle and electrostatics. We then computed cosine similarities between each surface feature in the generated structures and the experimental structure. Lastly, ProteinMPNN[28] and ESM1v[29] scores were calculated. ProteinMPNN is a deep learning-based sequence design method that can decode amino acid sequences from structural representations of proteins. ProteinMPNN can also be used to generate a log-likelihood score for any given sequence. Wimley–White is a measure of residue hydrophobicity, which was applied to surface residues in this case using SURFMAP.

An additional set of filters was created to narrow down the final set of sequences. Sequences were required to be in the top 75th percentile for both ProteinMPNN and ESM1v scores, sequences were filtered on length to exclude sequences that were too short, a conservative pLDDT filter of 90 was used and an acceptable range for net charge of the proteins was established. After this, sequences were selected manually in an attempt to cover sequence identities in the range of 90–97% to the entire HyPB sequence with high-quality sequences. During this manual selection process, sequences with a higher proportion of the key residues were selected for and any sequences that had particularly bad scores in any of the calculated metrics were avoided. A final selection of 22 sequences was made.

In silico deep mutational scan

ESM1v was used in a zero-shot version where the Poecliopsis amino acid sequence was given as an input. ESM1v creates a fitness score for all possible amino acids for residue position by calculating a log odds ratio, assuming an additive model when multiple substitutions exist. Then, the sum is made over the substituted positions and the sequence is masked at every substituted position[29].

Variant prediction was run in Google Colab Pro with one A-100 GPU with 80 GB of RAM. The script used to run the variant prediction can be found on GitHub (https://github.com/Alejo945/IS-HyPB). The output is a TSV file with all possible variants and their scores.

Plasmid DNA sequences

Transposase ORF amino acid sequences were codon-optimized for *Homo sapiens* and ordered and synthesized as gene fragments to TWIST biosciences. Gene fragments were cloned into a cytomegalovirus-based expression vector by Golden Gate assembly using Esp3I restriction enzyme. Transposon (cargo vector) plasmid sequences were defined as the first 150 bp from the transposon ends from both 5′ and 3′ TIR sequences and synthesized as gene fragments by TWIST biosciences with added overhangs for golden gate assembly. An EF1α RFP poly(A) expression cassette was included between the TIR. Triple mutant (×3, R372A;K375A;D450N in *Trichoplusia ni*) residue selection was performed by aligning the ortholog sequences to the *T. ni PiggyBac* mutated sequence. All plasmid sequences are available in Supplementary Table 1.

Cell culture

Hek293T cells (Invitrogen, R70007), were cultured in DMEM supplemented with high glucose (Gibco, Thermo Fisher), 10% FBS, 2 mM glutamine, 100 U per ml penicillin and 0.1 mg ml^{-1} streptomycin at 37 °C in a 5% CO$_2$ incubator.

PCR excision activity assay

To detect excision in bioprospected transposases, 120,000 cells were seeded per adherent p24 well 1 day before transfection. Plasmid DNA was mixed at a 1:3 ratio of transposase and RFP transposon, with 0.035 pmol of transposase used per p24 well plate. Then, 48 h after transfection, cells were collected and plasmid extraction was performed using an NZYMiniprep kit (NZYtech, MB01001). TIR-flanking primers (Supplementary Table 4) were used to detect transposon excision. The 2,900-bp and 1,200-bp bands indicated nonexcised and excised transposon, respectively.

Nontargeted transposon integration fluorescence assay

To evaluate stable transposon integration activity, 120,000 cells were seeded per adherent p24 well a day before transfection. Plasmid DNA was mixed with and RFP transposon at a ratio of 1:3:5, with 0.035 pmol of transposase used per p24 well plate. For transfection experiments, cells were transfected with polyethyleneimine (PEI, Thermo Fisher Scientific) at a 1:3 ratio of DNA and PEI in Opti-MEM. RFP expression of the transposon cargo vector was assessed 2 days and 20 days after transfection using cell cytometry with the Cytek Aurora CS system. The RFP signal at day 20 was considered indicative of stable transgene integration.

Transposon excision fluorescence assay

To quantify the excision activity of AI-generated transposases, a fluorescent excision reporter system was used. HEK293T cells were seeded in 24-well plates at a density of 120,000 cells per well 24 h before transfection to ensure approximately 70% confluency on the day of transfection. Transfections were performed in 24-well plates using PEI (Thermo Fisher Scientific) at a 1:3 ratio of DNA and PEI in Opti-MEM (Thermo Fisher). Transposase-expressing plasmid was cotransfected with plasmid containing a disrupted mCherry reporter sequence flanked by transposase recognition sites, leading to mCherry restoration upon excision (Supplementary Fig. 6). Transposase and transposon plasmids were mixed at a 1:3 ratio, with a total of 0.035 pmol of transposase. Then, 72 h after transfection, cells were collected and mCherry reporter expression was assessed by flow cytometry using the Cytek Aurora CS system.

Targeted transposon integration digital PCR assay

To quantify targeted integration of AI-generated transposases in the FiCAT system, C2C12 cells (American Type Cell Collection, CRL-1772) were cultured in DMEM (Gibco, Thermo Fisher) supplemented with 10% FBS, 2 mM L-glutamine, 100 U per ml penicillin and 0.1 mg ml^{-1} streptomycin. Cells were maintained in a 37 °C incubator with 5% CO$_2$. Electroporation was conducted using the E Cell Line 4D-Nucleofector X Kit S (Lonza). On the day of electroporation, cells were washed with PBS, detached using trypsin–EDTA (Gibco) and adjusted to a concentration of 2×10^5 cells per condition. The cell suspension was prepared in 20 µl of nucleofection master mix buffer, consisting of

16.4 µl Nucleofector solution and 3.6 µl of supplement 1 (Lonza). Subsequently, each condition was conucleofected with a DNA plasmid encoding the triple-mutant variants (PB×3), Cas9, different guide RNAs (gRNAs) and transposon plasmids in a 1:1:3:3 molar ratio, using a maximum of 10% of the final sample volume. Lastly, each condition was transferred into Nucleocuvette vessels and electroporation was carried out using the CD-137 program. After electroporation, 100 µl of prewarmed complete medium was added and cells were carefully resuspended and transferred into a 24-well plate containing 500 µl of complete medium for recovery and expansion. Then, 4 days after electroporation, the cells were processed as follows: (1) one third were collected for genomic extraction; (2) one third were analyzed for GFP reporter expression by flow cytometry using the Cytek Aurora CS system; and (3) one third were maintained in culture until episomal disappearance. Genomic extraction was performed using Qiagen DNeasy blood and tissue kit. Primers and probes were obtained from PrimeTime qPCR probes (Integrated DNA Technologies). The assay was designed using an endogenous control and evaluating the junction PCR for both integration orientations. Reaction mixtures (44 µl) were prepared containing QIAcuityDx Universal master mix (1×), $MgCl_2$ (6.28 mM), primers (0.73 µM), probes (0.63 µM), a restriction enzyme (0.25 U per µL) and 12.5 ng of sample DNA. These mixtures were loaded onto a QIAcuityDx Nanoplate 26k 24-well (260001) for quantification, following the preparation protocol provided in the QIAcuityDx Universal master mix kit (260102). Thermal cycling protocol consisted of an initial enzyme activation step at 95 °C for 2 min, followed by 40 cycles of a two-step amplification: denaturation at 95 °C for 15 s and annealing and extension at 60 °C for 30 s. For digital PCR analysis, the absolute DNA quantification per sample (copies per genome) was determined using QIAcuity Software. Primer sequences are described in Supplementary Table 6.

Targeted transposon integration fluorescence and qPCR assay

To quantify targeted integration of bioprospected transposases in the FiCAT system, Plasmids encoding the triple-mutant variants (PB×3) were cotransfected with Cas9, gRNA AAVS1-3, transposase and transposon plasmids at a 1:1:3:5 molar ratio in 0.5 M Hek23T cells seeded in a p6 plate the day before transfection. Cells were analyzed for RFP expression 2 days after transfection to estimate transfection efficiency using cell cytometry with the Cytek Aurora CS system. Cells were maintained in culture to measure overall integration levels after 3 weeks. In parallel, to enrich cells for junction qPCR, two rounds of enrichment by GFP sorting were conducted with BD FACSAria (Biosciences), 1 week and 2 weeks after transfection. Genomic DNA was extracted using Quiagen DNeasy blood and tissue kit column 4 days after the second sorting. A 3′ junction PCR was performed and sequenced on an Illumina MiSeq Nano kit 500 cycles (v2). A 3′ junction qPCR was performed to compare targeted integration across bioprospected transposases.

Targeted transposon integration GFP reconstitution assay

To quantify targeted integration in AI-generated *PiggyBac* transposases in the FiCAT system, a previously described GFP reconstitution assay[52] was used. For GFP targeted integration assays, a reporter HEK293T cell line containing genomically integrated 2/2 GFP was transfected using a 1/2 GFP encoding transposon (Supplementary Fig. 6). A total of 240,000 2/2 GFP HEK293T reporter cells were seeded in a 12-well plate 1 day before transfection. Cells were transfected with Lipofectamine 3000 (Invitrogen, L3000001) using Cas9, 2/2 GFP-targeting gRNA, transposase and transposon plasmids at a 1:1:3:5 molar ratio. Cells were analyzed for GFP expression 5 days after transfection to estimate targeted integration efficiency using cell cytometry with the Cytek Aurora CS system. The 2/2 GFP was integrated using the *Sleeping Beauty* (SB100x) transposase system[53]. Reporter DNA sequences are available in supplementary Table 3.

Nontargeted transposon integration fluorescence assay in T cells

To assess nontargeted integration of the *PiggyBac* and AI-generated orthologs in T cells, peripheral blood mononuclear cells from two different donors, isolated from buffy coats and cryopreserved, were thawed and seeded on p24-coated plates containing anti-CD3/CD28 (1:1,000; BD Sciences) at a density of 1 × 10^6 cells per ml in 3 ml of CTS OpTmizer T cell expansion SFM medium (Thermo Fisher), supplemented with interleukin (IL)-7 and IL-15 (10 ng ml^{-1} each; Miltenyi Biotec). Buffy coats were obtained from the Barcelona Blood and Tissue Bank upon institutional review board approval.

For nontargeted integration in bioprospected orthologs, on the third day of culture, electroporation was conducted using the P3 primary cell 4D-Nucleofector X kit (Lonza). Cells were washed with PBS (Capricorn) and adjusted to a concentration of 7.5 × 10^5 cells per condition. The cell suspension was prepared in 20 µl of nucleofection buffer, consisting of 16.4 µl of P3 primary cell Nucleofector solution and 3.6 µl of supplement 1 (Lonza). Subsequently, 1 µg of each DNA plasmid was added to the suspension and electroporation was carried out using the EO-115 nucleofection program. The minimal backbone GenCircle-TIR_CAR19-GFP transposon plasmid was used (GenCircle, manufactured by Genscript). For each evaluated transposase, conditions with transposase + transposon and transposon only were electroporated in duplicates to differentiate between episomal and integrated signals. Following electroporation, 80 µl of complete medium was added and cells were incubated at 37 °C for 20 min. The cells were then carefully resuspended and transferred to a fresh p24 plate containing 500 µl of medium for recovery and expansion. Approximately one third of the well volume was used for flow cytometric analysis using the Aurora system (Cytek) to assess RFP expression levels at 4 and 7 days after transfection.

For nontargeted integration of AI-generated orthologs, On the third day of culture, electroporation was conducted using the P3 primary cell 4D-Nucleofector X kit (Lonza). Cells were washed with PBS (Capricorn) and adjusted to a concentration of 1 × 10^6 cells per condition. The cell suspension was prepared in 20 µl of nucleofection buffer, consisting of 16.4 µl of P3 primary cell Nucleofector solution and 3.6 µl of supplement 1 (Lonza). Subsequently, 1 µg of each DNA plasmid was added to the suspension and electroporation was carried out using the EH-115 nucleofection program. The minimal backbone GenCircle-TIR_CAR19-GFP transposon plasmid was used (GenCircle, manufactured by Genscript).For each evaluated transposase, conditions with transposase + transposon and transposon only were electroporated in duplicates to differentiate between episomal and integrated signals. Following electroporation, 80 µl of complete medium was added and cells were incubated at 37 °C for 20 min. The cells were then carefully resuspended and transferred to a fresh p24 plate containing 500 µl of medium for recovery and expansion. Medium supplemented with H-151 (MedChemExpress, HY-112693) STING inhibitor at 2 µM was added. Approximately one third of the well volume was used for flow cytometric analysis using the Aurora system (Cytek) to assess GFP expression levels at 4 and 7 days after transfection.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Experimentally tested transposon sequence files are available in Supplementary Table 1. Top active transposon and transposase plasmids were deposited to Addgene.

Code availability

Model fine-tuning and *PiggyBac* generation code is available from Github (https://github.com/Integra-tx/Piggybac_bioprospecting_pipeline).

References

35. Kitts, P. A. et al. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* **44**, D73–D80 (2016).
36. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
37. Krause, G. R., Shands, W. & Wheeler, T. J.Sensitive and error-tolerant annotation of protein-coding DNA with BATH. *Bioinform. Adv.* **4**, vbae088 (2024).
38. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
39. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
40. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
41. Lu, S. et al. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265–D268 (2020).
42. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
43. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
44. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
45. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
46. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
47. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
48. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
49. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
50. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
51. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
52. Pallarès-Masmitjà, M. et al. Find and cut-and-transfer (FiCAT) mammalian genome engineering. *Nat. Commun.* **12**, 7071 (2021).
53. Mátés, L. et al. Molecular evolution of a novel hyperactive *Sleeping Beauty* transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.* **41**, 753–761 (2009).

Acknowledgements

We thank C. Feschotte for feedback on the transposon bioprospecting pipeline, G. M. Church for advice on approaches for machine learning, AI and directed evolution and M. Artigues for her help in designing the initial filtering process for the Progen2 generated sequences. Integra Therapeutics received funding from NEOTEC (CDTI, SNEO-20222363). Universitat Pompeu Fabra received funding from UPGRADE (European Union Horizon 2020, grant agreement no. 825825) and Ministerio de Economia, Industria y Competitividad de España (Plan Estatal 2013-2016 (grant agreement no. PID2020-118597RB-I00).

Author contributions

D.I., A.S.M. and M. Güell conceptualized the study. A.A. and D.I. designed the bioprospecting pipeline. A.A. implemented the bioprospecting pipeline. J.L.-V. implemented the LLM and fine-tuning work with help from N.F., A.A. and D.I. D.I. and J.J.-W. designed the experiments with help from R.D. and M. Gallo. M. Gallo, R.D. and J.J.W. performed the cell experiments. I.H. assisted with the sequence assembly. A.R. and P.P. performed the T cell work. F.B. contributed to genome data accession and zero-shot modeling. J.H.-V. performed insertional profiling and molecular characterization of editing and transposition outcomes. M.S.-G. analyzed the targeted integration data. D.I. and M. Güell supervised the study. A.A., D.I. and J.V.L. plotted the data. A.A., D.I., M. Güell and J.L.V. wrote the paper with contributions from all authors.

Competing interests

A.A., J.L.-V., J.J.-W., M. Gallo, M.S.-G., R.D., M. Güell, A.S.-M., N.F. and D.I. are employed or have consulted for Integra Therapeutics. M. Güell and A.S.-M. are shareholders of Integra therapeutics. D.I., M. Güell, A.S.-M., A.A. and R.D. have filed a patent application (US Patent application no. 63/505485) related to this work.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-025-02816-4.

Correspondence and requests for materials should be addressed to Dimitrije Ivančić or Marc Güell.

Peer review information *Nature Biotechnology* thanks Zoltán Ivics, Jesse Owens and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature portfolio

Corresponding author(s): Marc Güell, Dimitrije Ivančić

Last updated by author(s): Jul 16, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collected from the public online database Dfam and NCBI. Pubmed. Experimental data (included in the paper)
Data analysis	Alphafold server, ITOL v7. IQ-TREE v.1.6.12, MUSCLE 5.1, UCSF Chimera 1.16, ProGen2, SURFMAP 2.1.0, PyMOL 3.0.4, TMAlign: 20220412 Associated code for analysis has been made available at https://github.com/Integra-tx/Piggybac_bioprospecting_pipeline.git, Additional codes used are clearly mentioned an cited in the repository.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

We have included a data availability statement in the manuscript.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender

Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used.
Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected.
Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.

Reporting on race, ethnicity, or other socially relevant groupings

Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status).
Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.)
Please provide details about how you controlled for confounding variables in your analyses.

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

Ethics oversight

Identify the organization(s) that approved the study protocol.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Each experiment with cell lines was performed with replicates, otherwise indicated.
Data exclusions	No data was excluded from the analysis
Replication	Experiments wih cell lines were performed in replicates, and each replicate had at least 2 technical replicates per condition.
Randomization	Samples allocations to experimental groups was random
Blinding	Blinding was not performed

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).

Research sample

State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.

Sampling strategy

Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a *rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*

Data collection

Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.

Timing

Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Non-participation

State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.

Randomization

If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.

Research sample

Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.

Sampling strategy

Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

Data collection

Describe the data collection procedure, including who recorded the data and how.

Timing and spatial scale

Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Reproducibility

Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.

Randomization

Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

Blinding

Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work? □ Yes □ No

Field work, collection and transport

Field conditions

Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).

Location

State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).

Access & import/export

Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).

Disturbance

Describe any disturbance caused by the study and how it was minimized.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern
☒	□	Plants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
□	☒	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Hek293T cell line (ATCC CRL-3216), C2C12 cell line (ATCC CRL-1772), T cells obtained from the Catalan public procurement system (Banc de Sang i teixits)
Authentication	All cell lines were purchased with authentication certificate and they were not authenticated after purchase
Mycoplasma contamination	All cell llines tested negative for mycoplasma contamination. Mycoplasma test was performed every 3 months.
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

□ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	*For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were*

Wild animals

caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Reporting on sex

Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	□	Public health
☒	□	National security
☒	□	Crops and/or livestock
☒	□	Ecosystems
☒	□	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☒	□	Demonstrate how to render a vaccine ineffective
☒	□	Confer resistance to therapeutically useful antibiotics or antiviral agents
☒	□	Enhance the virulence of a pathogen or render a nonpathogen virulent
☒	□	Increase transmissibility of a pathogen
☒	□	Alter the host range of a pathogen
☒	□	Enable evasion of diagnostic/detection modalities
☒	□	Enable the weaponization of a biological agent or toxin
☒	□	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.

ChIP-seq

Data deposition

□ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

□ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Transfected cell cultures with DAPI staining

Instrument

BD LSR Fortessa; BD Biosciences. Blue 488nm laser with 530/30 filter and Yellow Green 561nm laser with 610/20 filte

Software

D FACSDiva version 6.2 and version 8.0.2

Cell population abundance	Purity after sorting was checked from the sorting population making sure it was higher than 90% for the library experiments were low % of cells were positive. Far Cytometry analyses more than 10,000 alive cells were analysed.
Gating strategy	Morphological related parameters (SSC-A vs. FSC-A) were used to exclude debris by P1 region. Subsequently P2 region (FSC-H vs FSC-A) and P3 region (DAPI vs FSC-A) were used to exclude aggregates and dead cells respectively. P4 region to isolate GFP population (GFP vs Autofluorescence using FITC and PerCP-Cy5-5-A lasers). P5 region was used to isolate and RFP expressing cells (RFP vs vs Autofluorescence using PE-Texas Red-A and PerCP-Cy5-5-A lasers).

□ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI □ Used □ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*
Volume censoring	*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings	*Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*
Effect(s) tested	*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: □ Whole brain □ ROI-based □ Both

Statistic type for inference (See Eklund et al. 2016)	*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*
Correction	*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
□	□ Functional and/or effective connectivity
□	□ Graph analysis
□	□ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	*Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*
Graph analysis	*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*
Multivariate modeling and predictive analysis	*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*