JCB Journal of Cell Biology

TOOLS

FREEDA: An automated computational pipeline guides experimental testing of protein innovation

Damian Dudka[1], R. Brian Akins[1], and Michael A. Lampson[1]

Cell biologists typically focus on conserved regions of a protein, overlooking innovations that can shape its function over evolutionary time. Computational analyses can reveal potential innovations by detecting statistical signatures of positive selection that lead to rapid accumulation of beneficial mutations. However, these approaches are not easily accessible to non-specialists, limiting their use in cell biology. Here, we present an automated computational pipeline FREEDA that provides a simple graphical user interface requiring only a gene name; integrates widely used molecular evolution tools to detect positive selection in rodents, primates, carnivores, birds, and flies; and maps results onto protein structures predicted by AlphaFold. Applying FREEDA to >100 centromere proteins, we find statistical evidence of positive selection within loops and turns of ancient domains, suggesting innovation of essential functions. As a proof-of-principle experiment, we show innovation in centromere binding of mouse CENP-O. Overall, we provide an accessible computational tool to guide cell biology research and apply it to experimentally demonstrate functional innovation.

Introduction

Purifying selection eliminates deleterious non-synonymous mutations, leading to conservation of amino acid sequence. In contrast, positive selection results in the accumulation of non-synonymous mutations that lead to functional innovation and adaptation (reviewed in Nielsen et al., 2005). Compelling examples of how positive selection has regulated protein function come from studying host–pathogen genetic conflicts. In these evolutionary arms races, positive selection leads to rapid accumulation of mutations in both viral proteins that help infect the host and host proteins that help evade the infection (reviewed in Daugherty and Malik, 2012; Sironi et al., 2015). To experimentally test functional innovation, evolutionary biologists swap protein regions (or entire alleles) from closely related species that are suspected to have diverged due to positive selection. This approach generates an "evolutionary mismatch" between the divergent protein and the cellular environment, revealing which protein function might have evolved adaptively (reviewed in Brand and Levine, 2021). For example, swapping a region of the TRIM5 protein between human and rhesus monkey suggested that positive selection shaped its role in fighting species-specific retroviral infections (Sawyer et al., 2005; Stremlau et al., 2005; Yap et al., 2005). Remarkably, variation at even single residues under positive selection can lead to functional changes, as in the human MAVS (Mitochondrial Antiviral Signaling) protein that has evolved to evade infection with hepaciviruses (Patel et al., 2012). These examples illustrate that innovation-guided functional analyses can complement more traditional conservation-guided approaches in revealing regulation of protein function.

Genetic conflicts, like those between host and pathogen, can result in recurrently changing selection pressure and recurrent adaptation of proteins regulating essential cellular processes. For example, pressure to maintain genome integrity at fertilization is thought to fuel a sexual conflict between paternal proteins that adapt to maximize the chance of fertilizing the egg and maternal proteins that adapt to prevent entry of more than one sperm (reviewed in Carlisle and Swanson, 2020). Similarly, selfish genetic elements such as transposons constantly disrupt genome integrity, leading to intragenomic conflicts and recurrent adaptation of DNA packaging proteins (reviewed in Brand and Levine, 2021). Centromere DNA sequences have also been proposed to act as selfish elements, raising the possibility of intragenomic conflict with centromere-associated proteins. Centromeres are repetitive DNA regions that direct chromosome segregation in mitosis and meiosis by assembling kinetochores, multiprotein structures that connect to spindle microtubules. Despite their essential function, centromeric DNA and proteins evolve rapidly across taxa, suggesting an evolutionary pressure to recurrently innovate. The centromere drive hypothesis proposes that selfish centromeric DNA sequences achieve non-

[1]Department of Biology, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA.

Correspondence to Michael A. Lampson: lampson@sas.upenn.edu; Damian Dudka: damiandudka0@gmail.com.

Mendelian segregation during asymmetric female meiosis, increasing their transmission to the egg. Fitness costs imposed by this selfish behavior would lead to recurrent adaptation of centromeric proteins to suppress the costs (Henikoff et al., 2001). While there is experimental evidence for selfish centromeric DNA, the impact of positive selection on centromeric protein function remains largely untested (reviewed in Dudka and Lampson, 2022).

The scarcity of experimental studies of adaptive evolution in centromeric proteins, in contrast to our increasingly detailed understanding of their conserved functions (Kixmoeller et al., 2020; McKinley and Cheeseman, 2016; Mellone and Fachinetti, 2021), reflects the general focus of cell biology research on protein conservation rather than innovation. This discrepancy is due in part to challenges in designing experiments to infer functional consequences of positive selection, but also to the complexity of methods needed to distinguish positive selection from neutral evolution of protein-coding sequences (reviewed in Anisimova and Liberles, 2012). A widely used method calculates the rate ratio of non-synonymous (dN) to synonymous (dS) substitutions per codon (dN/dS ratio; Goldman and Yang, 1994; Kimura, 1977; Muse and Gaut, 1994) using multiple sequence alignment of closely related orthologs, which are homologous genes that arise when speciation occurs. This approach assumes that synonymous mutations are neutral, while deleterious non-synonymous mutations are purged by purifying selection. An enrichment of non-synonymous relative to synonymous substitutions within the alignment suggests recurrent adaptation to a constantly changing selection pressure (types of recurrent evolution are discussed in Maeso et al., 2012). Well-established computational suites such as PAML (Phylogenetic Analysis by Maximum Likelihood; Yang, 2007) and HyPhy (Hypothesis Testing using Phylogenies; Pond et al., 2005) offer a number of tools that can reliably detect statistical signatures of positive selection but are seldom used by cell biologists because expertise in computational biology and molecular evolution is required to generate the input data, and the output is rarely provided in an intuitive visual format.

Automated molecular evolution pipelines that incorporate the abovementioned tools have been developed (see Fig. S1 for a non-exhaustive list), but their complexity and the need for user-provided input still render them inaccessible to experimental cell biologists with limited computational skills. Increasing this access requires a "one-click" application that (1) offers a simple graphical user interface, (2) fully automates input preparation, (3) finds orthologs despite the lack of genomic annotations, (4) reduces parameterization, and (5) provides intuitive visual representation of the output. Here, we present FREEDA (Finder of Rapidly Evolving Exons in Diverse Assemblies), a fully automated, end-to-end pipeline designed for cell biologists seeking to apply an evolutionary lens by testing for statistical evidence of positive selection in their favorite proteins. FREEDA provides the key functionalities listed above, including a unique ability to map positively selected residues onto any predicted protein structure. As a proof-of-principle, we first use FREEDA to map positive selection across centromeric proteins in rodents, as mice are currently the only experimentally tractable cell biological model system for centromere drive (reviewed in Dudka and Lampson, 2022). Guided by these computational analyses, we use the evolutionary mismatch approach to provide experimental evidence of functional innovation in the centromeric protein CENP-O.

Results

Overview of the FREEDA pipeline

FREEDA is a stand-alone application with an intuitive graphical user interface (GUI) operating on UNIX systems (MacOS and Linux; Windows users, please see documentation). An overview and documentation of the pipeline are provided at https://ddudka9.github.io/freeda/ with a more detailed walkthrough in Materials and methods. FREEDA first downloads the reference genome of the user-selected species and prepares input data for the gene of interest by connecting to genomic, protein, and protein structure databases (Fig. 1; blue). Next, FREEDA downloads a preselected set of non-annotated genome assemblies related to the reference species, performs a BLAST (Basic Local Alignment Search Tool) search for orthologs of the gene of interest, and uses the reference species data to find orthologous sequences (Fig. 1; orange). Combining several well-established molecular evolution tools, FREEDA aligns all coding sequences, builds phylogenetic trees, determines the likelihood that positive selection has shaped the evolution of the gene, and estimates the probability that given residues have evolved under positive selection (Fig. 1; brown). Key results are displayed within the GUI (Fig. 2 A) and all results are saved into the "Results-current-date" folder generated in a location selected by the user ("Set directory"; Fig. 2 A). These files include the BLAST output, nucleotide alignment, phylogenetic tree, protein alignment, residue mapping onto reference coding sequence, and residue mapping onto protein structure. The raw data and intermediate alignment files are saved in the "Raw_data" folder. Since FREEDA finds orthologs by downloading entire genomic assemblies, the user is advised to select an external data storage device (e.g., a hard drive) when setting the directory. A stable internet connection is also required to allow communication with various databases (Fig. 1).

Advantages over existing automated pipelines

Several features distinguish FREEDA from currently available automated pipelines (Fig. S1). First, FREEDA is fully automated, requiring only a gene name, and distributed as a self-contained application that does not require installation or compilation of any additional programs, except for a straightforward installation of the widely used protein structure viewer PyMOL (The PyMOL Molecular Graphics System, Version 2.0, Schrödinger, LLC) for MacOS users. Second, FREEDA uses a defined set of non-annotated genomic assemblies that ensure high statistical power of the analysis while absolving the users from manually curating their input. As new genomic assemblies become available, they will be incorporated into new FREEDA releases. Third, FREEDA automatically maps residues with the highest probabilities of having evolved under positive selection onto protein structure models by querying the AlphaFold database

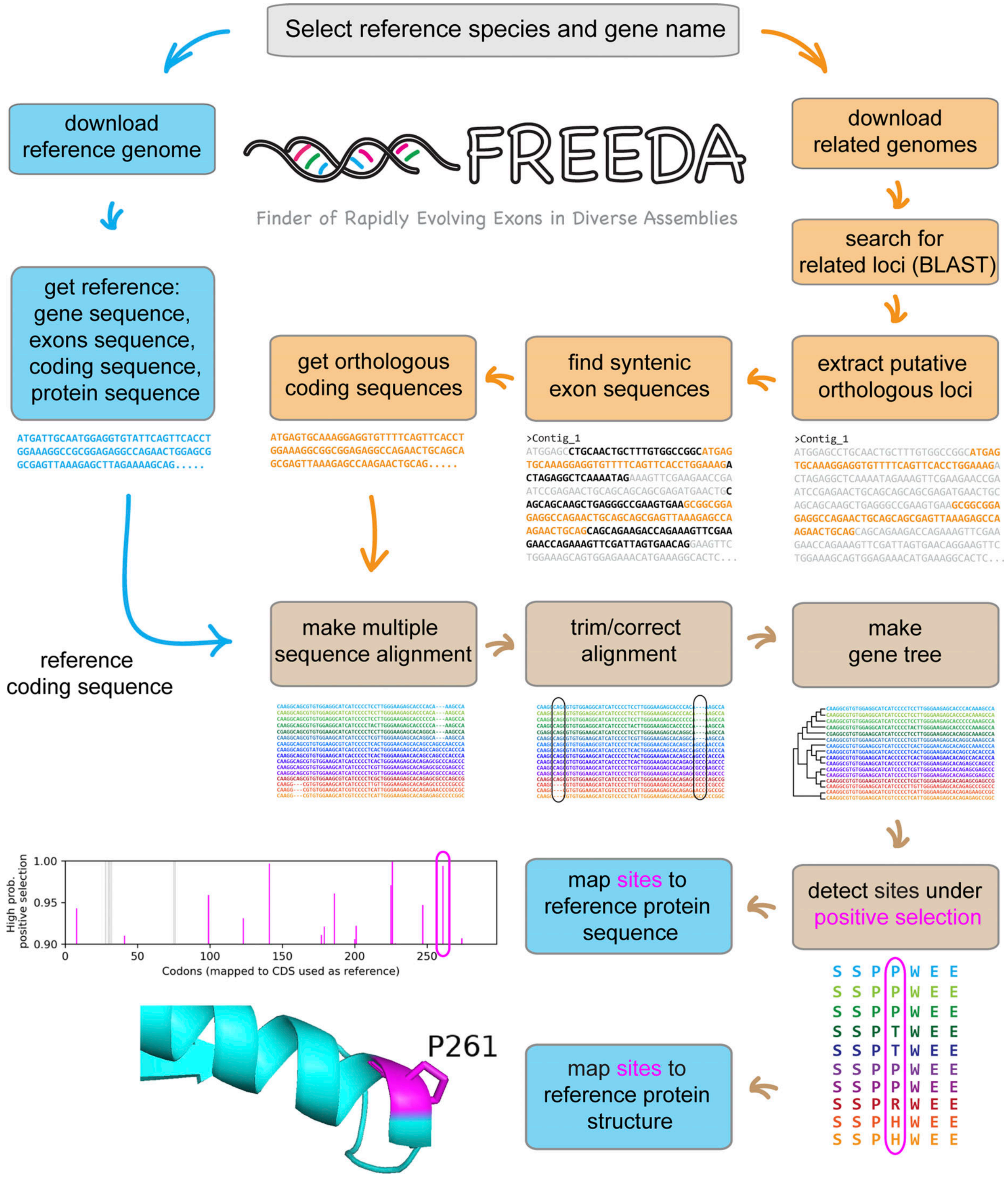
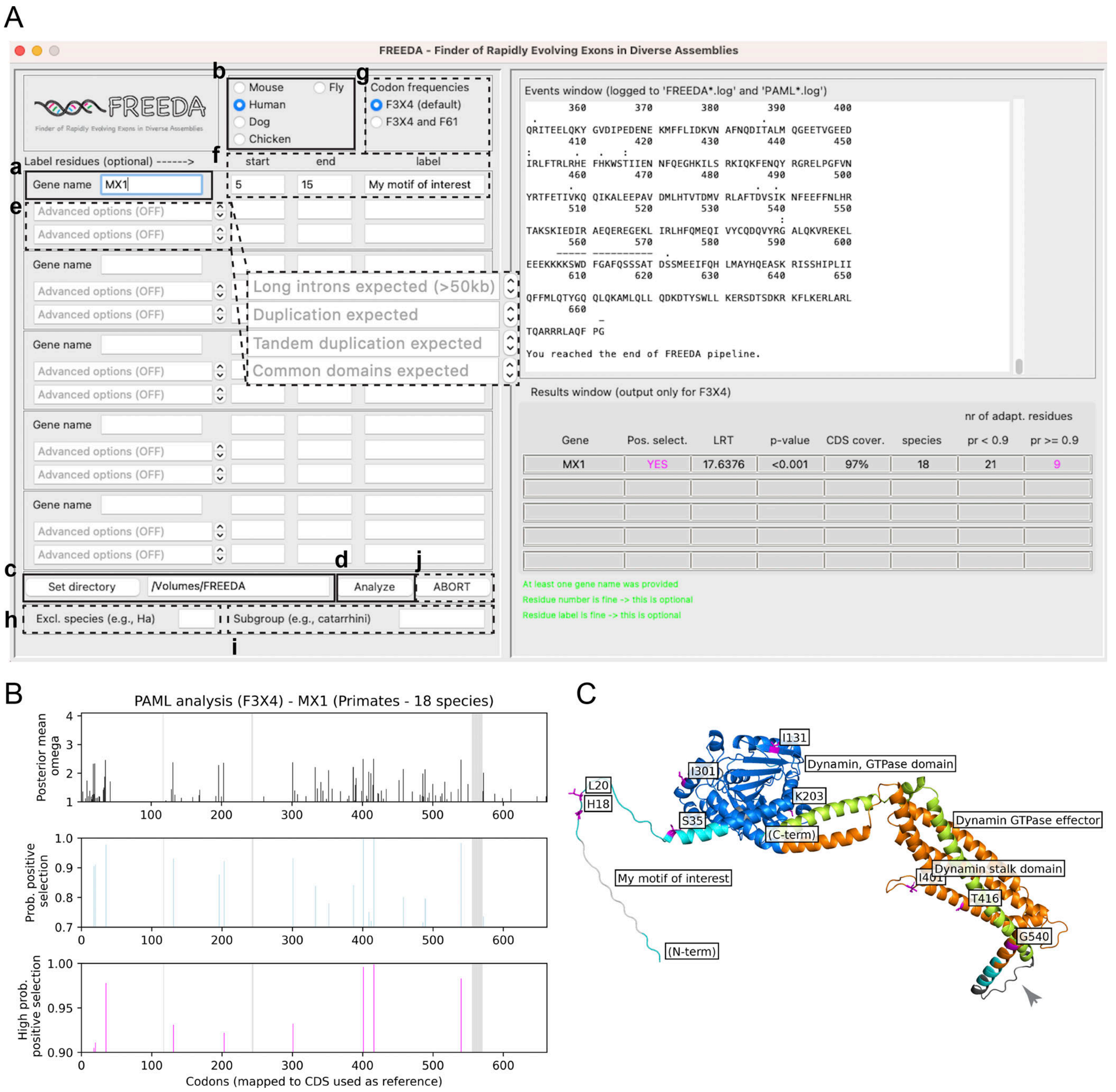

Select reference species and gene name
download reference genome
FREEDA
Finder of Rapidly Evolving Exons in Diverse Assemblies
download related genomes
search for related loci (BLAST)
get reference: gene sequence, exons sequence, coding sequence, protein sequence
get orthologous coding sequences
find syntenic exon sequences
extract putative orthologous loci
ATGATTGCAATGGAGGTGTATTCAGTTCACCT
GGAAAGGCCGCGGAGAGGCCAGAACTGGAGCG
GCGAGTTAAAGAGCTTAGAAAAGCAG.....
ATGAGTGCAAAGGAGGTGTTTTCAGTTCACCT
GGAAAGGCGGCGGAGAGGCCAGAACTGCAGCA
GCGAGTTAAAGAGCCAAGAACTGCAG.....
>Contig_1
ATGGAGCCTGCAACTGCTTTGTGGCCGGCATGAG
TGCAAAGGAGGTGTTTTCAGTTCACCTGGAAAGA
CTAGAGGCTCAAAATAGAAAGTTCGAAGAACCGA
ATCCGAGAACTGCAGCAGCAGCGAGATGAACTGC
AGCAGCAAGCTGAGGGCCGAAGTGAAGCGGCGGA
GAGGCCAGAACTGCAGCAGCGAGTTAAAGAGCCA
AGAACTGCAGCAGCAGAAGACCAGAAAGTTCGAA
GAACCAGAAAGTTCGATTAGTGAACAGGAAGTTC
TGGAAAGCAGTGGAGAAACATGAAAGGCACTC...
>Contig_1
ATGGAGCCTGCAACTGCTTTGTGGCCGGCATGAG
TGCAAAGGAGGTGTTTTCAGTTCACCTGGAAAGA
CTAGAGGCTCAAAATAGAAAGTTCGAAGAACCGA
ATCCGAGAACTGCAGCAGCAGCGAGATGAACTGC
AGCAGCAAGCTGAGGGCCGAAGTGAAGCGGCGGA
GAGGCCAGAACTGCAGCAGCGAGTTAAAGAGCCA
AGAACTGCAGCAGCAGAAGACCAGAAAGTTCGAA
GAACCAGAAAGTTCGATTAGTGAACAGGAAGTTC
TGGAAAGCAGTGGAGAAACATGAAAGGCACTC...
reference coding sequence
make multiple sequence alignment
trim/correct alignment
make gene tree
High prob. positive selection
1.00
0.95
0.90
0
50
100
150
200
250
Codons (mapped to CDS used as reference)
map sites to reference protein sequence
detect sites under positive selection
S S P P W E E
S S P P W E E
S S P P W E E
S S P T W E E
S S P T W E E
S S P P W E E
S S P P W E E
S S P R W E E
S S P H W E E
S S P H W E E
P261
map sites to reference protein structure

Figure 1. **Overview of the FREEDA pipeline.** The schematic shows the main steps, with more details in Materials and methods. Launching the FREEDA application opens a graphical user interface (gray), which prompts selection of a reference species and a gene name. First, operating on the selected reference species (blue path), FREEDA downloads the genome and uses it to curate input for the gene of interest (protein sequence, exon sequences, coding sequence, and gene sequence). Second, operating on closely related species (orange path), FREEDA downloads non-annotated genomes, searches for putative orthologous loci, retrieves these loci, finds syntenic (homologous to the reference locus) exons, and assembles coding sequences of orthologous genes based on the intron–exon boundaries known for the reference gene. Third, operating on the multiple coding sequences (brown path), FREEDA makes and curates a multiple sequence alignment, generates a phylogenetic gene tree, and detects statistical signatures of positive selection using established models measuring the rate ratio of non-synonymous to synonymous substitutions. Fourth, FREEDA maps sites with the highest probability of positive selection onto both the reference coding sequence and the structure prediction of the reference protein.


A
FREEDA - Finder of Rapidly Evolving Exons in Diverse Assemblies
FREEDA
Finder of Rapidly Evolving Exons in Diverse Assemblies
a
b
c
d
e
f
g
h
i
j
Label residues (optional) ------>
Advanced options (OFF)
Long introns expected (>50kb)
Duplication expected
Tandem duplication expected
Common domains expected
Set directory
Analyze
ABORT
Excl. species (e.g., Ha)
Subgroup (e.g., catarrhini)
B
PAML analysis (F3X4) - MX1 (Primates - 18 species)
Posterior mean omega
Prob. positive selection
High prob. positive selection
Codons (mapped to CDS used as reference)
C
I131
Dynamin, GTPase domain
I301
K203
L20
H18
S35
(C-term)
Dynamin GTPase effector
Dynamin stalk domain
I401
T416
G540
My motif of interest
(N-term)

Gene name MX1

○ Mouse ● Human ○ Dog ○ Chicken ○ Fly

Codon frequencies ● F3X4 (default) ○ F3X4 and F61

start	end	label
5	15	My motif of interest

Set directory /Volumes/FREEDA

Events window (logged to 'FREEDA*.log' and 'PAML*.log')

```
       360        370        380        390        400
 .                                  .
QRITEELQKY GVDIPEDENE KMFFLIDKVN AFNQDITALM QGEETVGEED
       410        420        430        440        450
:        .   .   :
IRLFTRLRHE FHKWSTIIEN NFQEGHKILS RKIQKFENQY RGRELPGFVN
       460        470        480        490        500
       .                                 .  .
YRTFETIVKQ QIKALEEPAV DMLHTVTDMV RLAFTDVSIK NFEEFFNLHR
       510        520        530        540        550
                                             :
TAKSKIEDIR AEQEREGEKL IRLHFQMEQI VYCQDQVYRG ALQKVREKEL
       560        570        580        590        600
     ----- ----------  .
EEEKKKKSWD FGAFQSSSAT DSSMEEIFQH LMAYHQEASK RISSHIPLII
       610        620        630        640        650

QFFMLQTYGQ QLQKAMLQLL QDKDTYSWLL KERSDTSDKR KFLKERLARL
       660
                _
TQARRRLAQF PG

You reached the end of FREEDA pipeline.
```

Results window (output only for F3X4)

						nr of adapt. residues	
Gene	Pos. select.	LRT	p-value	CDS cover.	species	pr < 0.9	pr >= 0.9
MX1	YES	17.6376	<0.001	97%	18	21	9

At least one gene name was provided
Residue number is fine -> this is optional
Residue label is fine -> this is optional

Figure 2. **Example analysis of the primate *MX1* gene. (A)** FREEDA's graphical user interface is divided into an input window (left half) and an output window (right half). The input window is used to provide a gene name (a), select the reference species (b), indicate where to save the data (c), and start the analysis (d). Optionally, the user can select advanced features ("Duplication expected," "Tandem duplication expected," "Long introns expected (>50 kb)," and "Common domains expected"; see online documentation: https://ddudka9.github.io/freeda/; e), label up to three regions of choice on the protein structure (f), select an additional codon frequency model (g), exclude species from the analysis (h), select a subgroup (i), and abort the analysis (j). The output window shows current tasks ("Events window," top) and key results for each gene ("Results window," bottom). The bottom part of the output window (green font) displays interactive messages guiding the user on how to provide input. **(B)** Putative adaptive sites are mapped onto the reference coding sequence. Graphs show recurrently changing residues (top, black bars), residues that are likely to have evolved under positive selection (middle, blue bars, probability ≥ 0.7), and most likely targets of positive selection (bottom, magenta bars, probability ≥ 0.9). Gray bars in all graphs show residues removed from the analysis. **(C)** Residues with the highest probability of positive selection (magenta) are mapped on the structural prediction model of the MxA protein (encoded by *MX1*) from AlphaFold. The N- and C-termini and known domains are automatically annotated, in addition to the user-specified region ("My motif of interest"). Regions removed from the analysis (arrowhead) are colored dark gray. To clearly show residues under positive selection, labels were modified manually in PyMOL (raw output is shown in Fig. S2 E).

containing structure predictions for nearly all known proteins (Jumper et al., 2021). Finally, by providing a simple GUI, FREEDA minimizes complexity compared to currently available pipelines, while offering a restrained number of advanced options (Fig. 2 A). Therefore, the user may consider FREEDA as an entry point to performing the first molecular evolution analyses of their proteins of interest.

FREEDA validation: Finding orthologous genes

To demonstrate that FREEDA's simplicity does not compromise its functionality, we first tested its ability to find orthologous sequences in non-annotated assemblies—a notoriously challenging task in the field. To remain unbiased, we randomly selected five genes from reference assemblies of rodents (*Murinae*; mouse), primates (*Simiiformes*; human), carnivores (*Carnivora*; dog), and birds (*Phasianidae*; chicken) and compared the FREEDA-identified orthologs to their annotations in 26 species available in the highly curated Ensembl database (Cunningham et al., 2022). While only 26/74 species used by FREEDA for these clades have Ensembl-annotated assemblies, we managed to analyze >120 orthologs. We confirmed the identities of all the FREEDA-identified orthologs by showing that they share (1) genomic location with Ensembl-annotated flanking genes and (2) an average of 99.9% (without insertions/deletions [indels]) or 90.3% (with indels) nucleotide sequence identity with Ensembl-annotated orthologs. The use of alternative exons and start codons explains the vast majority of sequence differences when indels are not excluded (see Table S1). These unbiased analyses validate FREEDA's ability to reliably detect orthologous genes.

To ensure rigor in detecting orthologs, we tested if FREEDA can distinguish them from paralogs, which form by a duplication event and may evolve under different selective pressures. FREEDA demonstrated the ability to resolve ancient duplications by correctly distinguishing *HERC5* orthologs from *HERC6* paralogs present within a dataset of previously curated human immune genes used to validate the DGINN pipeline (Detect Genetic INNovations; Picard et al., 2020; see raw data of the analyzed dataset in additional online supplemental material: https://doi.org/10.5281/zenodo.7997737). Additionally, we tested if FREEDA could resolve tandem duplications (duplicated genes located side by side) and retro-duplications (intron-less mRNA that was reverse-transcribed and inserted back into the genome). Using the "Tandem duplication expected" option (see GUI; Fig. 2 A), FREEDA successfully distinguished primate genes *H4C1* and *H4C2*, both encoding histone H4 and located merely 5 kb apart with 85% nucleotide sequence identity (additional supplementary materials). Visual examination of the nucleotide alignment revealed that one *H4C2* ortholog (in *Plecturocebus donacophilus*) lost the ancestral start codon and likely pseudogenized. In such cases, the user may choose to rerun the analysis using the "Exclude species" option (see GUI; Fig. 2 A). We further used the "Duplication expected" option (see GUI; Fig. 2 A) to show that FREEDA could correctly distinguish *KIF4A*, encoding a kinesin motor, from its retroduplicate *KIF4B*. These genes are an example of a recent duplication that occurred in a common ancestor of primates, retaining 96% nucleotide sequence identity (Florio et al., 2018; additional supplementary materials). We also found that *KIF4B*, and not *KIF4A*, has likely evolved under positive selection, suggesting that a duplication event spurred adaptive evolution of this kinesin. While human KIF4A regulates chromosome segregation (Mazumdar et al., 2004), cellular transport (Peretti et al., 2000), and anti-viral response (Gad et al., 2022), KIF4B remains poorly studied. Together, these analyses demonstrate that FREEDA reliably finds orthologous sequences even when a gene has undergone duplication.

FREEDA validation: Detecting statistical signatures of positive selection

We then tested if FREEDA could accurately detect statistical signatures of positive selection in genes with known evolutionary histories. To do so, we used a dataset of 23 primate (*Simiiformes*) genes whose statistical signatures of positive selection (or lack thereof) have been previously defined. The dataset included 19 genes curated to validate the DGINN pipeline (Picard et al., 2020). Analyzing a set of 19 primate species, FREEDA found 18 orthologs with 98% coding sequence coverage (median values; Table S2). Consistent with the literature, FREEDA found statistical signatures of positive selection in *TRIM5*, *MAVS*, *SAMHD1*, *IFI16*, *ZC3HAV1*, *RSAD2*, GBP5, *MX1*, *APOBEC3F*, and *NBN* (Table S2). Although previous studies also reported that positive selection has likely shaped the evolution of *BST2* (using nine primate species; Gupta et al., 2009; van der Lee et al., 2017), FREEDA only found a weak statistical signature of positive selection in that gene (P = 0.0864; Table S2), which likely stems from the lineage of New World monkeys (Liu et al., 2010). Of six genes whose evolutionary history is less clear, with results dependent on the method used (Picard et al., 2020), FREEDA found statistical evidence of positive selection in only one (*SERINC3*; Table S2), highlighting the stringency of the analysis. Of six genes whose adaptive evolution has been previously deemed unlikely, FREEDA detected a signature of positive selection in one, *TREX1*, a nuclease that guards genome integrity (Picard et al., 2020; Table S2). One of the residues with the highest probability of positive selection (serine at position 166 in human; probability = 0.97; Table S2) is proximal to a primate-specific DNA-binding site (arginine at position 164 in humans; Zhou et al., 2022), suggesting that adaptive evolution has shaped DNA recognition. Consistent with our finding, divergent DNA binding sites in *TREX1* regulate DNA recognition (Zhou et al., 2022). We suspect that differences in regions removed from the analysis are responsible for the disparity between published results (Picard et al., 2020) and ours. Altogether, using previously curated datasets allowed us to objectively validate our pipeline and provided additional insight into the evolutionary history of these genes.

To further validate accuracy of the pipeline at the level of single residues, we compared specific sites that have likely evolved under positive selection found by FREEDA to those previously mapped in *MAVS*, *MX1*, *SAMHD1*, and *TRIM5* (Laguette et al., 2012; Lim et al., 2012; Mitchell et al., 2012; Patel et al., 2012; Sawyer et al., 2005; van der Lee et al., 2017). Exact matching of probabilities for each residue was not expected due to differences in algorithms for aligning orthologous sequences (see Materials and methods for details). Nevertheless, FREEDA found

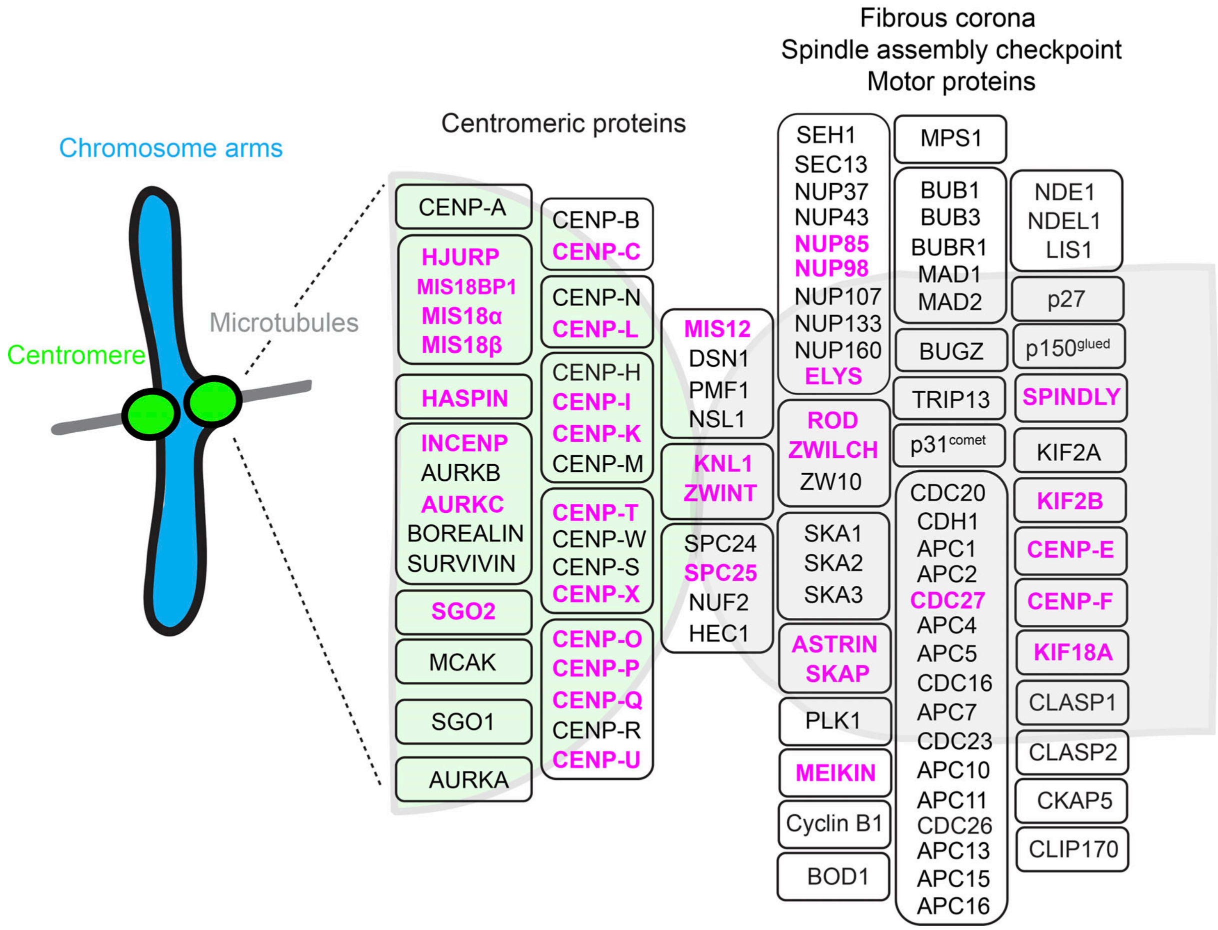


Figure 3. **Positive selection across the rodent centromere.** Centromere-associated proteins are grouped to reflect their approximate positions between chromatin and microtubules. Relative protein placement is informed by binding partners, but not all known interactions are depicted. Proteins forming complexes are grouped together. Magenta: proteins that have likely evolved under positive selection.

statistical signatures of positive selection in all published sites, except for those located in regions removed from the alignment to ensure its high quality (five residues in each *SAMHD1* and *MX1*; two residues in *TRIM5*; Fig. S2, A–E; and Table S3). Using *MX1* as an example (Fig. 2 A), FREEDA maps detected sites onto the reference coding sequence (Fig. 2 B) and onto structural prediction models generated by AlphaFold (Fig. 2 C; Jumper et al., 2021). Overall, these analyses demonstrate that FREEDA can retrieve expected sites with previously reported signatures of positive selection and showcase FREEDA's key results visualization features.

Finally, we tested if FREEDA can reliably detect statistical signatures of positive selection in rodent genomes (*Murinae*). As a test dataset, we selected 104 centromeric genes, 42 of which have been previously analyzed using a smaller number of species (up to 11; Kumon et al., 2021). Analyzing a set of 19 *Murinae* species, FREEDA found 16 orthologs with 94% coding sequence coverage (median values; Table S4). Consistent with our previous findings of pervasive evolutionary innovation across the rodent centromere (Kumon et al., 2021), FREEDA found that 36/104 genes have likely evolved under positive selection (Fig. 3 and Table S4). Corroborating our previous results, FREEDA detected statistical signatures of positive selection in genes encoding CENP-C, CENP-I, CENP-T, HJURP, INCENP, MIS18BP1, KNL1, and SGO2. In contrast, DSN1 and HEC1 did not show statistical signatures of positive selection. This discrepancy is likely due to a difference in coding sequence coverage between the analyses (higher in FREEDA) or it reflects a higher statistical power due to more orthologs (up to 19 used by FREEDA), which facilitates distinguishing between positive selection and relaxation of purifying selection. This high statistical power revealed several previously unknown targets of positive selection, including components of the fibrous corona, which helps capture microtubules (CENP-F, SPINDLY, ZWILCH, ROD, NUP85, NUP98, and ELYS; reviewed in Kops and Gassmann, 2020), microtubule motors (CENP-E, KIF2B, and KIF18A), and protein kinases (AURKC and HASPIN). To further validate our findings, we repeated the analyses with rat as reference species. Since the quality of the available rat genome annotation is lower than that of mouse, FREEDA was able to collect reliable input data for only 89/104 genes. As expected, we found statistical evidence (or lack thereof) of positive selection in almost exactly the same genes as when using mouse as reference (85/89 genes; see Discussion; Table S4). Overall, these tests show that despite its simplicity for the user, FREEDA is a fully functional and dependable tool to detect statistical signatures of positive selection.

Using FREEDA to derive evolution-guided hypotheses

To test if FREEDA can help derive evolution-guided hypotheses, we leveraged its ability to map residues that have likely evolved under positive selection onto protein structures. We found statistical evidence of positive selection within ancient (retained across long evolutionary timescales) protein domains of centromeric proteins, suggesting that adaptive evolution shaped essential protein functions (Fig. 4, A–F). For instance, we detected residues with high probability of having evolved under positive selection in the ancient Yippee domain (Roxström-Lindquist and Faye, 2001) of MIS18β (encoded by *Oip5* in mouse), which participates in centromere chromatin assembly (reviewed in Zasadzińska and Foltz, 2017). In addition to its divergent N- and C-termini, one of the most likely adaptive residues (arginine at position 76 in mouse; probability = 0.98) is located within the loop-forming CXXC motif of the Yippee domain (Fig. 4, A and B), which is required for MIS18 complex assembly at centromeres (Fujita et al., 2007; Stellfox et al., 2016; Subramanian et al., 2016).

Similarly, we found strong statistical evidence of positive selection in one of the loops of an ancient protein kinase domain (reviewed in Taylor and Kornev, 2011) in the meiosis-specific Aurora kinase C (AURKC, asparagine at position 150 in mouse, probability = 0.98; Fig. 4, C and D), which helps correct erroneous kinetochore-microtubule attachments (Balboula and Schindler, 2014). In contrast, we found no recurrent changes in the related AURKA and AURKB kinases (Fig. S3, A–C). These data suggest that positive selection has uniquely tuned the kinase activity of the specialized meiotic Aurora kinase, consistent with previous reports of adaptive evolution of reproduction genes (Jagadeeshan and Singh, 2005; Nielsen et al., 2005; Swanson and Vacquier, 2002).

Finally, we found statistical evidence that positive selection shaped evolution of the ancient double RWD domain (RING-WD-DEAD; Tromer et al., 2019) of CENP-O, which regulates kinetochore–microtubule attachments by forming the CENP-OPQUR complex (Amaro et al., 2010; Bancroft et al., 2015; Chen et al., 2021; Hori et al., 2008; Singh et al., 2021; Fig. 4, E and F). RWD domains are prevalent structural modules that facilitate protein–protein interactions across the centromere (Schmitzberger and Harrison, 2012; Tromer et al., 2019). CENP-O shares a high structural similarity with its binding partner CENP-P, which also shows statistical signatures of positive selection within its double RWD domain (Fig. S4, A and B). Furthermore, some of the residues that have evolved under positive selection with the highest probability are located in or near loops and turns flanking highly structured α-helices and β-sheets in CENP-O and -P C-terminal RWD domains (Fig. 4 E and Fig. S4 A). Based on these results, we propose that positive selection has regulated essential functions of centromeric proteins by acting on loops and turns of ancient domains, consistent with previous reports of frequent innovation of flexible regions in other proteins (Afanasyeva et al., 2018; Nilsson et al., 2011; Ridout et al., 2010). Altogether, we demonstrate that FREEDA can help derive evolution-guided hypotheses by highlighting protein domains whose function has likely been shaped by adaptive evolution.

Using FREEDA to infer molecular mechanisms regulated by positive selection

Each of the proteins discussed above (MIS18β, AURKC, and CENP-OP) functions as part of a complex. To infer mechanisms regulated by positive selection in this context, we aligned FREEDA-annotated protein structure predictions of mouse proteins (Fig. 4) to experimentally solved structures of their orthologs in complex with binding partners (see Materials and methods for details). Two loops formed by CXXC motifs within the Yippee domain of MIS18β together give rise to a tetrahedral module whose four conserved cysteines bind a zinc ion (Subramanian et al., 2016), likely stabilizing protein conformation (Nguyen et al., 2020). Aligning mouse MIS18β to the crystal structure of the fission yeast MIS18 Yippee-like domain (Subramanian et al., 2016; Fig. 5, A and B) shows that the side chain of arginine at the positively selected position 76 in mouse likely faces the opening of the tetrahedral module. This finding is consistent with XX residues regulating the function of CXXC motifs in other proteins (Quan et al., 2007). Alternatively, R76 could mediate MIS18α and MIS18β heterodimerization (Subramanian et al., 2016). Therefore, we hypothesize that positive selection favored amino acid changes within the CXXC motif to modulate MIS18 complex stability. Consistent with functional innovation of CXXC motifs, we also found recurrently changing residues within the second CXXC motif of MIS18β (glycine at position 135 in mouse; Fig. 5, A and B) and in the first CXXC motif of its binding partner MIS18α (serine at position 57 in mouse; Fig. S5, A–C), albeit the probability that they have evolved under positive selection was lower (probabilities = 0.88 and 0.77, respectively). These data suggest that positive selection in the loops of the ancient Yippee domains regulated centromere assembly by modulating stability of the MIS18 complex.

AURKB and AURKC kinase activity requires binding to a conserved domain of INCENP (INner-CENtromere Protein; reviewed in Krenn and Musacchio, 2015). Aligning the *Mus musculus* (Mm) AURKC protein kinase domain and MmINCENP AURK-binding domain to the crystal structure of the orthologous human domains (Abdul Azeez et al., 2019) shows the side chain of positively selected asparagine at position 150 in mouse in close proximity to tyrosine at conserved position 827 in MmINCENP. This finding suggests modulation of INCENP binding and, therefore, kinase activity by positive selection (Fig. 5, C and D). The rodent AURKC activation loop also contains a recurrently changing, albeit less likely adaptive residue (serine at position 156 in mouse; probability = 0.77; Fig. 5, C and D) whose side chain reaches toward the AURKC ATP-binding site (marked by the inhibitor BRD-7880; Abdul Azeez et al., 2019; Fig. 5, C and D). These data suggest that positive selection in the loop of the ancient protein kinase domain of AURKC regulated meiotic functions by modulating kinase activity.

Double RWD domains mediate the formation of CENP-OP heterodimers, allowing recruitment of the CENP-OPQUR complex to centromeres (Pesenti et al., 2018; Schmitzberger and Harrison, 2012). Aligning the FREEDA-annotated CENP-O and -P C-terminal RWD domains to the experimentally solved human CENP-OPQUR complex (Yatskevich et al., 2022) suggests that positive selection shaped opposite sides of the CENP-OP

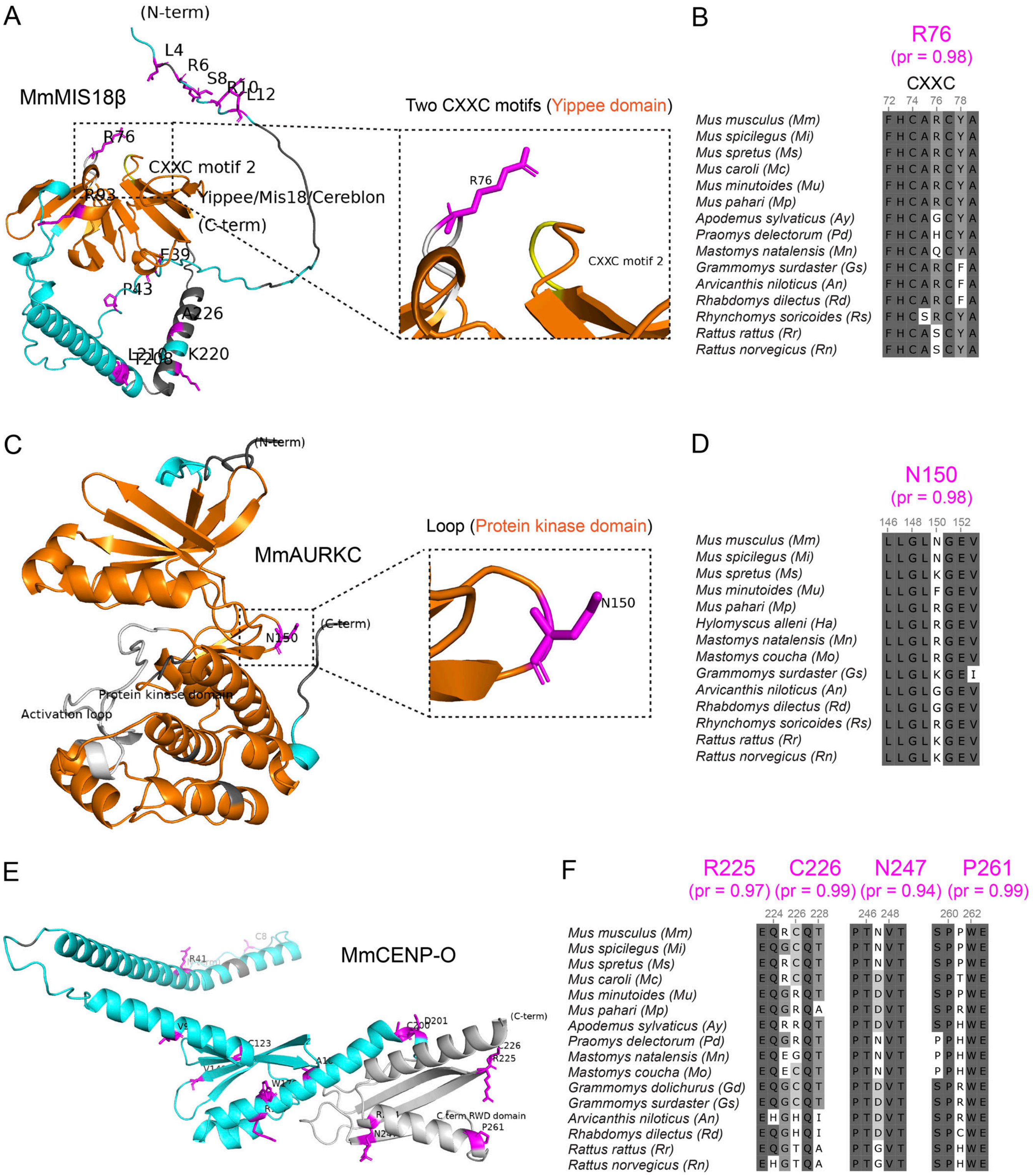

A
(N-term)
L4
R6
S8
R10
L12
MmMIS18β
R76
CXXC motif 2
R93
Yippee/Mis18/Cereblon
(C-term)
F39
P43
A226
L210
T208
K220
Two CXXC motifs (Yippee domain)
R76
CXXC motif 2
B
R76
(pr = 0.98)
CXXC
72 74 76 78
Mus musculus (Mm) F H C A R C Y A
Mus spicilegus (Mi) F H C A R C Y A
Mus spretus (Ms) F H C A R C Y A
Mus caroli (Mc) F H C A R C Y A
Mus minutoides (Mu) F H C A R C Y A
Mus pahari (Mp) F H C A R C Y A
Apodemus sylvaticus (Ay) F H C A G C Y A
Praomys delectorum (Pd) F H C A H C Y A
Mastomys natalensis (Mn) F H C A Q C Y A
Grammomys surdaster (Gs) F H C A R C F A
Arvicanthis niloticus (An) F H C A R C F A
Rhabdomys dilectus (Rd) F H C A R C F A
Rhynchomys soricoides (Rs) F H C S R C Y A
Rattus rattus (Rr) F H C A S C Y A
Rattus norvegicus (Rn) F H C A S C Y A
C
(N-term)
MmAURKC
N150
(C-term)
Protein kinase domain
Activation loop
Loop (Protein kinase domain)
N150
D
N150
(pr = 0.98)
146 148 150 152
Mus musculus (Mm) L L G L N G E V
Mus spicilegus (Mi) L L G L N G E V
Mus spretus (Ms) L L G L K G E V
Mus minutoides (Mu) L L G L F G E V
Mus pahari (Mp) L L G L R G E V
Hylomyscus alleni (Ha) L L G L R G E V
Mastomys natalensis (Mn) L L G L R G E V
Mastomys coucha (Mo) L L G L R G E V
Grammomys surdaster (Gs) L L G L K G E I
Arvicanthis niloticus (An) L L G L G G E V
Rhabdomys dilectus (Rd) L L G L G G E V
Rhynchomys soricoides (Rs) L L G L R G E V
Rattus rattus (Rr) L L G L K G E V
Rattus norvegicus (Rn) L L G L K G E V
E
MmCENP-O
C8
R41
V9
C123
A1
W1
D201
C200
(C-term)
C226
R225
C-term RWD domain
P261
N24
F
R225
(pr = 0.97)
C226
(pr = 0.99)
N247
(pr = 0.94)
P261
(pr = 0.99)
224 226 228 246 248 260 262
Mus musculus (Mm) E Q R C Q T P T N V T S P P W E
Mus spicilegus (Mi) E Q G C Q T P T N V T S P P W E
Mus spretus (Ms) E Q R C Q T P T N V T S P P W E
Mus caroli (Mc) E Q R C Q T P T D V T S P T W E
Mus minutoides (Mu) E Q G R Q T P T D V T S P P W E
Mus pahari (Mp) E Q G R Q A P T D V T S P R W E
Apodemus sylvaticus (Ay) E Q R R Q T P T D V T S P H W E
Praomys delectorum (Pd) E Q G R Q T P T N V T P P H W E
Mastomys natalensis (Mn) E Q E G Q T P T N V T P P H W E
Mastomys coucha (Mo) E Q E C Q T P T N V T P P H W E
Grammomys dolichurus (Gd) E Q G C Q T P T D V T S P R W E
Grammomys surdaster (Gs) E Q G C Q T P T D V T S P R W E
Arvicanthis niloticus (An) E H G H Q I P T D V T S P R W E
Rhabdomys dilectus (Rd) E Q G H Q I P T D V T S P C W E
Rattus rattus (Rr) E Q G T Q A P T G V T S P H W E
Rattus norvegicus (Rn) E H G T Q A P T G V T S P H W E

Figure 4. **Positive selection in loops and turns within ancient protein domains. (A, C, and E)** Ribbon diagrams show annotated structural prediction models of mouse proteins, generated automatically by FREEDA and visualized in PyMOL without manual modifications. Residues with the highest probability of having evolved under positive selection are colored magenta, and a subset of these is shown in snippets of the multiple sequence alignments in Murinae. **(B, D, and F)** Dark gray: highly conserved residues; gray: less conserved residues; white: non-synonymous substitutions. **(A and B)** MIS18β (MmMIS18β) shows the Yippee domain (orange) and two CXXC motifs (labeled by the user within the GUI, gray and yellow). The label for CXXC motif 1 is not visible to accommodate labeling of the R76 residue. The two CXXC motifs are enlarged with the R76 residue (magenta) within motif 1. **(C and D)** AURKC (MmAURKC) shows the protein kinase domain (orange) and activation loop (labeled by the user within the GUI, gray). A loop within the protein kinase domain is enlarged, with N150 shown in the multiple sequence alignment. **(E and F)** CENP-O (MmCENP-O) shows the C-terminal RWD domain (labeled by the user within the GUI, gray). The most likely adaptive residues of C-terminal RWD domain are shown in the multiple sequence alignment.

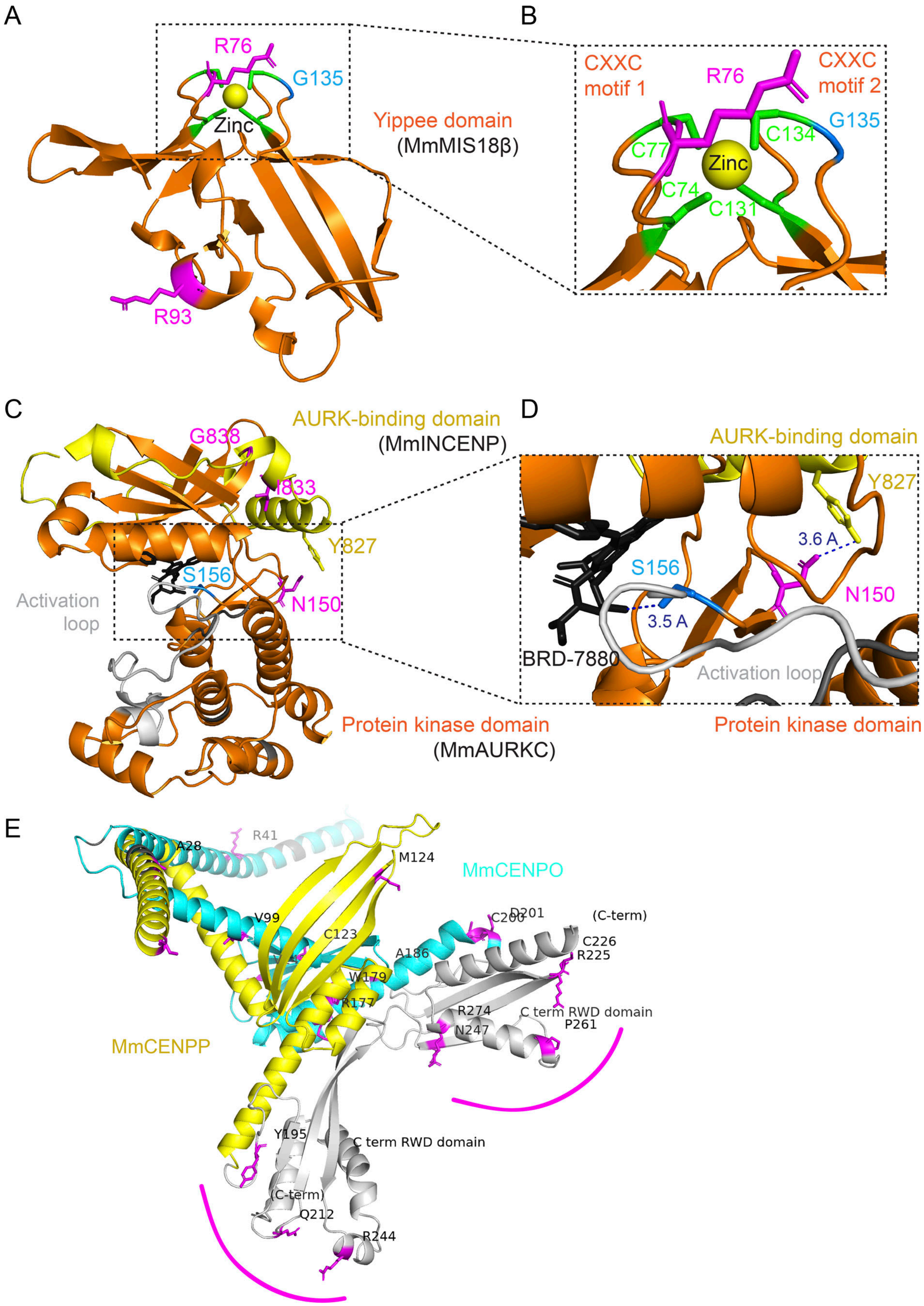


Figure 5. **Putative molecular interactions of likely adaptive residues. (A)** Yippee domain (orange) of MmMIS18β aligned to the Yippee-like domain of fission yeast MIS18 (PDB 5JH0; Subramanian et al., 2016). FREEDA automatically annotates residues with the highest probability of having evolved under

positive selection (probability ≥ 0.9, magenta). Manual annotations show residues with lower probability (probability ≥ 0.7, blue) and conserved cysteins (green). **(B)** Enlarged CXXC motifs forming a tetrahedral module holding a zinc ion. **(C)** MmAURKC protein kinase domain (orange, with activation loop in gray) and AURK-binding domain of mouse INCENP (MmINCENP; yellow), both aligned to the human AURKC-INCENP complex with the BRD-7880 inhibitor bound to the ATP-binding site (black; PDB 6GR8; Abdul Azeez et al., 2019). Annotations show likely adaptive residues (magenta, probability ≥ 0.9; blue, probability ≥ 0.7) and conserved residue Y827 of MmINCENP (yellow). **(D)** Enlarged ATP-binding site. Dashed lines show the closest distance from side chains of residues that likely evolved under positive selection (S156 or N150) to the BRD-7880 inhibitor or to Y827 of MmINCENP. **(E)** MmCENP-O (blue) and MmCENP-P (yellow) aligned to human CENP-O and -P (HsCENP-O and -P) from the human CENP-OPQUR complex (PDB 7PB8; Yatskevich et al., 2022). Gray: C-terminal RWD domains; magenta: the most likely adaptive residues (probability ≥ 0.9). Magenta arcs highlight most likely adaptive residues within the C-terminal RWD domains facing opposite sides of the heterodimer, which likely interface with other centromeric proteins.

heterodimer (Fig. 5 E) and therefore is unlikely to have impacted heterodimerization. In yeast, C-terminal RWD domains of CENP-O and -P orthologs bind to CENP-Q and -U orthologs to form the COMA complex (Hinshaw and Harrison, 2019; Schmitzberger and Harrison, 2012). We were unable to reliably align mouse CENP-Q and -U to the human CENP-OPQUR complex, likely due to long unstructured regions in CENP-Q and -U, but the striking pattern of likely adaptive residues in C-terminal RWD domains facing the outside of the heterodimer suggests that positive selection regulated binding to nearby centromeric components (Fig. 5 E). We find statistical signatures of positive selection in CENP-Q and -U in rodents (Fig. 3), suggesting that positive selection regulated interactions between CENP-OPQUR complex components. Altogether, these analyses of multiple centromere proteins demonstrate how FREEDA-annotated structures can be used to generate hypotheses for how positive selection might have regulated essential protein functions.

Experimental testing of functional protein innovation

To test our hypothesis that loops and turns in ancient protein domains regulate their essential functions, we chose to focus on CENP-O because FREEDA suggests that positive selection operated on residues flanking α-helices and β-sheets of both rodent and primate C-terminal RWD domains (Fig. S6, A–D), and centromere binding provides a straightforward functional assay. We used mouse oocytes for these experiments because they are an established model system for centromere drive, the most likely selective pressure sculpting evolution of centromeric proteins, and thus a natural context to probe for functional protein innovation. To create an evolutionary mismatch (see Introduction), we introduced GFP-tagged full-length mouse (control) or rat (divergent) CENP-O at similar expression levels (Fig. S7 A). Mouse CENP-O localized to centromeres as expected, but rat CENP-O was nearly undetectable at mouse centromeres (Fig. 6, A and B), indicating functional innovation in centromere binding. To test if the C-terminal RWD domain is responsible for that innovation, we compared three chimeric rat CENP-O constructs with different regions of mouse CENP-O: N-terminal (N-terminal tail and N-terminal helix), central (N-terminal RWD domain and central helix), or C-terminal (C-terminal RWD domain; Fig. 6, A and B; and Fig. S7 A). Only the mouse C-terminal RWD domain could rescue, albeit not fully, the localization of rat CENP-O to mouse centromeres. In an inverse experiment, a chimera of mouse CENP-O with the rat C-terminal RWD domain failed to localize to mouse centromeres (Fig. 6, C and D; and Fig. S7 B). Together, these results demonstrate that mouse-specific innovation in the C-terminal RWD domain is required for CENP-O binding to mouse centromeres. Within this domain, 10 out of 13 residues that differ between mouse and rat are putatively adaptive (probability ≥ 0.5; Fig. S8). Almost all (9/10) of these residues flank highly structured α-helices or β-sheets (±1 amino acid), consistent with our hypothesis that positive selection drives functional innovation of ancient domains in centromeric proteins by acting on their loops and turns. Swapping five of the most likely adaptive residues in the mouse C-terminal RWD domain to those found in rat did not, however, reduce centromere localization of mouse CENP-O (Fig. S9, A–C). Similarly, swapping equivalent rat residues within the C-terminal RWD domain of rat CENP-O to mouse-specific ones (in addition to six mutations in other parts of the protein) did not restore its centromere localization (Fig. S10, A–C). These analyses highlight the difficulty in attributing innovation to specific residues given the number of possible combinations as well as the potential for epistasis (Starr and Thornton, 2016). Altogether, we show that our fully automated molecular evolution pipeline can guide experimental testing of functional protein innovation.

Discussion

The motivation to develop FREEDA was to catalyze participation of the cell biology community in testing functional consequences of protein innovation. We demonstrate that detecting statistical signatures of positive selection, which implicates functional innovation, can be fully automated by compiling widely used bioinformatic and molecular evolution tools into a single pipeline (Fig. 1). FREEDA's simple and user-friendly GUI makes it a suitable entry point for experimentalists who may have limited programming skills (Fig. 2). Moreover, by leveraging the ever-growing pool of newly sequenced but not yet annotated genomic assemblies, FREEDA bypasses the requirement for obtaining tissue samples and cloning the genes of interest to have sufficient numbers of orthologs from closely related species to detect signatures of positive selection. Nevertheless, as with any fully automated tool, FREEDA has limitations. First, by inferring orthologs based on the annotated reference sequence, rather than experimentally validated transcripts, FREEDA does not account for tissue-specific splicing, shifts in intron–exon boundaries, or the use of alternative exons (see Table S1). Despite this caveat, using independently annotated rat sequences as reference led to the same result as using mouse annotations in 85/89 centromeric genes (note that relatively poorer rat genome annotation quality prevented reliable input generation for 15 genes; Table S4). Nevertheless, isoforms that substantially differ from the

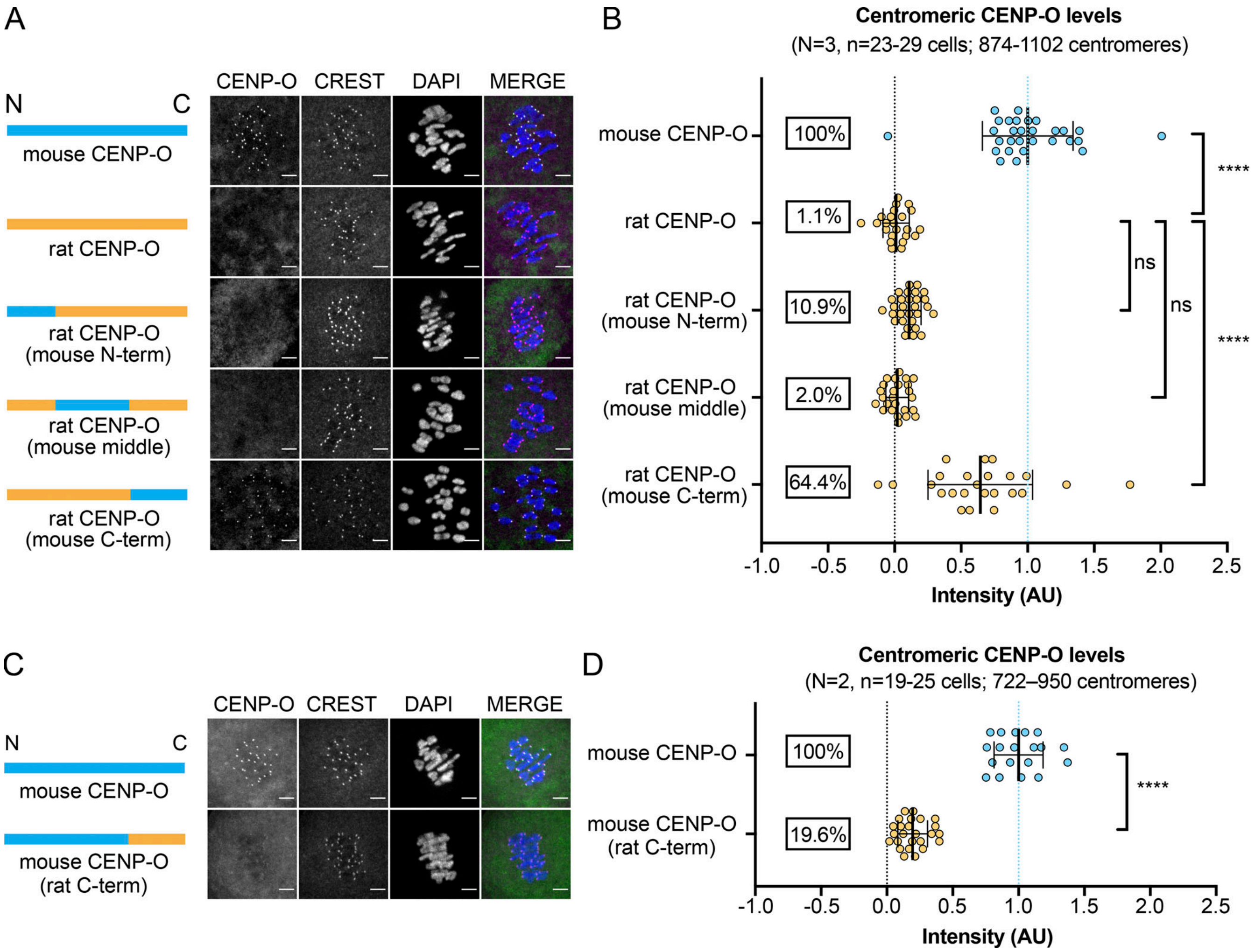


Figure 6. **Experimental evidence of functional innovation in the CENP-O C-terminal RWD domain.** Mouse oocytes expressing mouse, rat, or chimeric CENP-O–GFP were fixed in meiosis I and stained for centromeres (CREST) and DNA (DAPI). **(A and C)** Images show maximum intensity projections; scale bars, 5 µm. **(B and D)** Graphs show CENP-O–GFP intensity at centromeres; for each construct, $n \geq 722$ centromeres from ≥19 cells from three (B) or two (D) independent experiments. Each spot represents one cell; bars: mean intensities with standard deviation; ****P < 0.0001, ns: not significant, one-way ANOVA with Tukey's multiple comparison test (B) or two-tailed Student's *T* test (D).

reference coding sequence might interfere with the accurate detection of positive selection. As an example, annotated variants of the rat NUP37 nucleoporin substantially differ at their C-termini from the reference mouse NUP37 sequence, suggesting the use of alternative exons (additional supplementary materials), which likely led to inconsistent signals of positive selection (P = 0.510 with mouse as reference vs. P = 0.0499 with rat as reference; Table S4). Second, to prioritize computational speed and reduce output complexity, FREEDA does not test for possible recombination events known to increase the probability of false positives when recombination rates are high (Anisimova et al., 2003). While estimated recombination frequencies are lower in vertebrates and insects as compared to yeast and protozoa, we cannot exclude the influence of recombination events (Stapley et al., 2017; Wilfert et al., 2007). Third, while FREEDA can robustly resolve gene duplications present in the ancestor of a selected taxon (e.g., primates), caution is advised when analyzing lineage-specific genes. For example, primate *MICA* (MHC class I chain-related gene A) is known to have duplicated from *MICB* in the common ancestor of hominoids and Old World monkeys (Florio et al., 2018). Therefore, searching for *MICA* orthologs across the entire primate taxon yields *MICB* coding sequences in New World monkeys (additional supplementary materials). In case lineage-specificity is suspected, we suggest using the "Subgroup" option (currently supporting: *hominoidea*, *catarrhini*, *caniformia*, and *melanogaster* subgroup) and/or the "Exclude species" option (Fig. 2 A). Fourth, FREEDA is designed to test for signatures of recurring (pervasive) positive selection acting on the entire taxon (e.g., primates) rather than episodic selection that may have led to adaptation in a specific branch (e.g., hominoids). Nevertheless, using the aforementioned "Subgroup" or "Exclude species" options allows narrowing of the phylogenetic window if needed. Finally, mapping FREEDA's results onto protein structures is not yet fully supported for carnivores (*Carnivora*) and birds (*Phasianidae*).

Our analyses of genes with known evolutionary histories demonstrate that FREEDA is reliable. Building on this validation, we provide the most detailed characterization of signatures of positive selection at rodent centromeres to date. Consistent with previous analyses (Kumon et al., 2021), we infer pervasive evolutionary innovation in domains of centromeric proteins that do not directly touch DNA, such as RWD (Fig. 3). Therefore, our data support the idea that fitness costs of centromere drive are suppressed by innovation in protein–protein interactions (Kumon et al., 2021; Rosin and Mellone, 2016) as well as protein–DNA interactions (Henikoff et al., 2001; Malik et al., 2002; Vermaak et al., 2002). Furthermore, by mapping regions that are likely under positive selection onto protein structures, we derive a hypothesis that positive selection acting on loops and turns of ancient domains impacts essential protein functions. For example, recurrent amino acid changes within the loops formed by CXXC motifs could regulate MIS18 complex formation. Similarly, recurrent changes in loops of the ancient AURKC protein kinase domain could modulate kinetochore–microtubule attachment dynamics, specifically in meiosis I. Both centromere assembly and microtubule detachment (in meiosis I) represent mechanisms potentially hijacked by selfish centromeres (Akera et al., 2019; Henikoff et al., 2001; Rosin and Mellone, 2016; Wu et al., 2018). These analyses provide a starting point for future experiments probing the functional impacts of innovation, including testing whether positive selection reduces fitness costs associated with centromere drive (reviewed in Dudka and Lampson, 2022).

Previous experiments in fruit flies, using evolutionarily mismatches of the L1 loop within the ancient histone fold domain of Cid$^{CENP\text{-}A}$, suggested functional innovation in a DNA-binding domain of a centromere protein (Rosin and Mellone, 2016; Vermaak et al., 2002). Here, we propose that positive selection in CENP-O may have regulated centromere binding via recurrent changes in loops and turns within an ancient C-terminal RWD domain (Fig. 6), which does not interact with DNA (Pesenti et al., 2022; Yatskevich et al., 2022). While we are unable to pinpoint the exact combination of residues responsible for functional innovation, the observation that most recurrently changing residues are within regions that flank the highly structured α-helices or β-sheets of that domain supports our hypothesis. CENP-O is expected to dock the CENP-OPQUR complex to centromeres (Eskat et al., 2012; Pesenti et al., 2018), promoting kinetochore–microtubule attachment stability (Amaro et al., 2010; Bancroft et al., 2015; Chen et al., 2021; Hori et al., 2008; Singh et al., 2021). Therefore, we propose that innovation in the CENP-O C-terminal RWD domain modulated interactions with other centromeric components (possibly CENP-Q and -U) necessary to form a stable CENP-OPQUR complex at centromeres (Foltz et al., 2006; Hori et al., 2008; Kagawa et al., 2014; Minoshima et al., 2005; Okada et al., 2006; Pesenti et al., 2018), potentially stabilizing kinetochore–microtubules to counteract destabilizing activities associated with driving centromeres (Akera et al., 2019; Wu et al., 2018). Future work using centromere drive models (reviewed in Dudka and Lampson, 2022) will be needed to experimentally test this idea. Overall, we show how FREEDA can help derive evolutionary hypotheses and guide experimental testing of functional innovation, starting from just a gene name, making it a powerful tool for incorporating evolutionary analyses into cell biology research and generating new insights into essential cellular processes.

Materials and methods

Resources and datasets

FREEDA was written in Python and compiled into a stand-alone application using pyinstaller (https://pyinstaller.org/en/stable/). Core packages used for the compilation (Table S5) were installed using standard package managers: pip (https://pypi.org/project/pip/) and conda (https://docs.conda.io/en/latest/). All selected genomic assemblies and Ensembl releases used to generate datasets are listed in Table S5. All data were collected using desktop computers: iMac (late 2015) with MacOS Monterey 12.6 (8 GB RAM; 4 CPU cores; 2.8 GHz Quad-Core Intel Core i5) and iMac (2017) with MacOS Ventura 13.2 (16 GB RAM, 2 CPU cores; 2.3 GHz Intel Core i5), or a laptop MacBook Pro (mid-2014) with MacOS Mojave 10.14.6 (16 GB RAM, 2 CPU cores; 3 GHz Intel Core i7).

Input extraction and identification of potential orthologous exons

FREEDA can be downloaded from an open-source repository: https://github.com/DDudka9/freeda/releases. When running the app for the first time, MacOS users will be prompted to download PyMOL, which renders a 3D result of the FREEDA analysis. To run the pipeline, the user needs to provide at least one gene name, select a reference species, and select a location where all the data will be stored. At least 100 GB of storage space is needed to analyze a single vertebrate taxon (e.g., primates), 20 GB is sufficient to analyze only flies, and 500 GB is recommended to analyze all taxons. Optionally, the user can (1) specify the coordinates of residues or domains of interest that will be labeled on the protein structure prediction, (2) customize the BLAST search and ortholog finding (see below), (3) exclude selected species from the analysis, and (4) narrow the analysis to a specific subgroup (e.g., *catarrhini*). Advanced users can also specify the codon frequency model used (F3X4 or F3X4 and F61; see PAML manual for details: http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf). The pipeline starts with downloading the reference genome (using NCBI Datasets [Sayers et al., 2021]) and then retrieving all possible UniProt IDs (UniProt Consortium, 2021) for a protein encoded by the gene of interest and matching them to AlphaFold database entries (Jumper et al., 2021). Next, FREEDA extracts protein sequence and coding sequence (using pyensembl package: https://github.com/openvax/pyensembl) from the Ensembl database (Cunningham et al., 2022) and exon sequences and gene sequence (using pybedtools [Dale et al., 2011]) from the downloaded reference genome. Visualization of residues that have likely evolved under positive selection requires that the protein sequence of the structural prediction and the protein sequence from the Ensembl database are identical (tested using the Biopython package; Cock et al., 2009). If the proteins are not identical, FREEDA performs the analysis without mapping the residues onto structure predictions. The first run triggers downloading of the selected reference genome (e.g., human), followed by the genomes of closely related species (e.g., *Simiiformes*), and then building of local BLAST databases (using BLAST + applications; Camacho et al., 2009). FREEDA queries these databases to find genomic coordinates of putative orthologous regions using tblastn algorithms (default identity threshold is set at 60% [or 30% for *Drosophila*] but can be increased to 80% [or 60% for *Drosophila*] by selecting the advanced option "Common domains expected") and retrieves corresponding nucleotide sequences (using pybedtools) from downloaded related genomes. Overall, these features allow fully automated generation of the input data needed to find orthologous coding sequences in non-annotated genomes.

Finding orthologous exons

FREEDA performs a multiple sequence alignment of each region found during the BLAST search to the reference coding sequence, BLAST sequences stitched together, the genomic locus these sequences reside in (contig), and the reference gene sequence using MAFFT (Multiple Alignment using Fast Fourier Transform; Katoh and Standley, 2013). Regions aligning to both the coding sequence and the reference gene sequence are considered putative exons. To determine if the putative exon is syntenic (resides in a homologous locus), the flanking sequence is compared with the introns of the reference gene separately at 5′ and 3′ ends. An exon is considered syntenic if at least one of the flanking regions is at minimum 75% (60% for *Drosophila*) identical to the reference intron over a stretch of at least 50 bp (30 bp for *Drosophila*). The identity is calculated as the Hamming distance (the number of different bases in a pair-wise comparison of two aligned sequences; Hamming, 1950; divided by the sequence length). The putative exon is called as not syntenic and discarded from the analysis if none of the flanking regions reaches the identity threshold and the exon itself is <80% (70% for *Drosophila*) identical to the reference exon over a stretch of the first 30 bp. Since introns are generally less conserved than exons, when the identity of a flanking region is uncertain (66–75%; 50–60% for *Drosophila*), a longer sequence is compared. Lowered values for detecting synteny in *Drosophila* genomes are due to an observed high divergence of intragenic regions and high rates of indels within orthologous loci.

To increase stringency in detecting synteny and reduce uncertainty resolving segmental gene duplications, the user can additionally select an advanced option "Duplication expected" that penalizes any exon that is syntenic only at one end (e.g., recent segmental duplication whose introns have not yet diverged significantly). Segmental duplications that lead to duplicated genes residing next to each other (tandem duplications) will not only have similar flanking regions but might also be difficult to align if residing on the same contig. To ensure robust analysis of tandem duplications, the "Tandem duplication expected" option limits the flanking region of each blast hit (leading to smaller contigs), decreasing the chance of tandemly duplicated genes residing on the same contig. In addition, to avoid retroduplications (mRNA-derived gene duplications; Kaessmann et al., 2009), FREEDA always discards exons that are at least 80% (70% for *Drosophila*) identical to the reference exon but lack the flanking regions (intron-less).

To preserve intron–exon boundaries, each putative exon is given a number and directly aligned to a reference exon of the same number. Therefore, exons do not need to reside on the same contig to form a complete coding sequence, which is helpful when querying genomic assemblies with short contigs. Very small exons (microexons; reviewed in Ustianenko et al., 2017) shorter than 18 bp cannot be reliably aligned and are discarded from the analysis. If the same putative exon is found on different contigs (e.g., due to a duplication), the contig containing fewer putative exons is discarded. If both contigs carry the same number of putative exons (likely due to heterozygosity of the orthologous locus or a very recent duplication), these are compared to corresponding reference exons, and the contig with a higher overall identity of exons is considered orthologous. Rare cases of mistakes in aligning intron–exon boundaries may lead to indels, which are resolved at later stages (e.g., the entire codon is removed in case of a 1-nt indel). To allow manual review, all the above-mentioned steps are logged and all the intermediate alignments are saved as raw data ("Raw_data" folder).

Manual verification of detected orthologs

Genomic location and nucleotide sequence identity of >120 FREEDA-identified orthologs representing 20 randomly selected genes (five per clade: *Murinae*, *Simiiformes*, *Carnivora*, *Phasianidae*) were compared with their annotations found in the Ensembl database. Genomic location (contig number) logged in the "FREEDA-current-date.log" file was compared to that of the expected flanking genes that were also analyzed by FREEDA. Nucleotide sequence identity with Ensembl-annotated orthologous coding sequences was measured using pairwise alignment and MAFFT protocol designed to limit over-aligning errors (Katoh and Standley, 2016). This approach generates large indels in the alignment and facilitates detection of alternative exons and start codons. Each alignment was visually inspected without manual curation. See detailed analysis and commentary in Table S1.

Manual verification of known recent gene duplications

The ability to distinguish tandem duplication (*H4C1* from *H4C2*) and recent retro-duplication (*KIF4A* from *KIF4B*) was tested by manual BLAST (blastn) of the nucleotide sequence of each ortholog identified by FREEDA ("GENE_raw_nucleotide_alignment.fasta" file) against the primate NCBI gene database (*Simiiformes*; taxid: 314294). For each gene, an orthologous sequence was always the highest scoring hit (by similarity) as opposed to a paralogous sequence. Exceptions were sister species *Aotus nancymaae* and *Callithrix jacchus*, whose identified *KIF4B* coding sequences were more similar to *KIF4A* than *KIF4B*. However, all the *KIF4B* orthologs detected by FREEDA were intron-less, consistent with *KIF4B* being a primate-specific retro-duplication of *KIF4A* (Florio et al., 2018), which FREEDA called correctly. Therefore, we are confident that FREEDA identified *KIF4B* orthologs for all species and not *KIF4A* paralogs (additional supplementary materials).

Building the multiple sequence alignment and phylogenetic gene tree

Detection of recurring amino acid substitutions requires a gapless, in-frame multiple sequence alignment. To avoid large gaps (suggesting incomplete coding sequences), FREEDA first removes entire coding sequences that are shorter than 90% compared with the reference sequence. To ensure high-quality alignment of the remaining sequences, we tested the commonly used aligners: MUSCLE (Edgar, 2004), PRANK (Löytynoja and Goldman, 2005), MACSE (Ranwez et al., 2011), and MAFFT (Katoh et al., 2002). We decided to use a modified MAFFT protocol designed to limit over-aligning errors (Katoh and Standley, 2016) as we found it both fast and accurate. To curate the alignment, FREEDA removes insertions that are defined as regions missing in the reference coding sequence and deletes stop codons (including premature ones). At this point, coding sequences that are <69% (60% for *Drosophila*) identical to the reference sequence are discarded as likely too divergent to produce a reliable alignment (based on Jeffares et al., 2015; Sievers et al., 2011) or misaligned. Additionally, remaining small gaps (deletions) and ambiguously aligned codons are removed from the alignment (using Gblocks; Castresana, 2000; Talavera and Castresana, 2007). Note that MAFFT is not codon-aware, which allows aligning incomplete sequences (e.g., missing some of the exons or containing indels) that cannot be expected to have a length of multiplication of 3. Therefore, to ensure that the aligning process and alignment curation did not alter the open reading frame of the aligned sequences, FREEDA compares the identity of the translated reference sequence within the curated alignment to the original reference protein sequence from the Ensembl database. Once 100% identity is confirmed, FREEDA translates the orthologous sequences and checks if >10 contiguous non-synonymous substitutions compared with the reference sequence are present. FREEDA considers such rare cases as likely frameshift events and removes the entire sequences from the alignment. Final in-frame multiple nucleotide sequence alignment is then used to build a phylogenetic gene tree (using RAxML; Stamatakis, 2014), which guides the widely used CODEML program from the PAML suite (Yang, 2007) to infer the enrichment of recurrent non-synonymous substitutions suggestive of positive selection. We strongly urge the user to manually verify the final protein alignment ("Results-Current-Date/Results/Protein_alignments/GENE_protein_alignment.fasta" file), ensuring that there are no obvious misaligned regions before considering the results of the PAML analysis (e.g., using the free software Unipro UGENE; Okonechnikov et al., 2012). In case of apparent misalignments, we suggest simply rerunning the analysis using the "Exclude species" option (Fig. 2 A). An example of a misaligned sequence can be found in the documentation.

Detection of positive selection

To detect statistical signatures of positive selection, FREEDA relies on the rate ratio of non-synonymous (dN) to synonymous (dS) substitutions (dN/dS > 1 suggests positive selection). However, most genes contain conserved regions that evolve under purifying selection (dN/dS < 1), which usually decreases gene-wide dN/dS below 1. Therefore, to find specific regions that likely evolved under positive selection, FREEDA uses "site models" of the CODEML program that allow for varying dN/dS between different codons. Each model describes a set of parameters (including dN/dS per codon; for details, see the official PAML guide, http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf, or a beginners guide; Jeffares et al., 2015) and either allows for sites (codons) with a dN/dS ratio of >1 (signature of positive selection; M8 and M2a models) or not (null hypothesis; M7 and M1a models). Using a maximum likelihood approach, CODEML then fits the parameters estimated from the data to each model. Significantly more likely fit (based on the likelihood ratio test) to the model that allows for codons with dN/dS > 1 indicates the presence of sites that have likely evolved under positive selection. Bayesian statistics (Bayes Empirical Bayes) are then used to estimate probabilities of positive selection acting on specific codons. FREEDA outputs the key results of the CODEML analysis within the GUI's "Results window" (likelihood ratio test value for M7 vs. M8 comparison, P value, and number of codons with the highest probability to have evolved under positive selection). Additionally, the results of the M1a vs. M2a comparison and the identity of specific codons under positive selection are saved in an Excel sheet ("Results-Current-Date/Results/Results_sheet" folder).

Visualization of residues under positive selection

FREEDA maps the protein sequence of the reference species from the multiple sequence alignment to the expected reference protein sequence and, if appropriate, introduces gaps that represent residues excluded from the analysis ("GENE_protein_alignment.fasta" file). Based on that mapping, FREEDA provides both 2D and 3D visual representations of residues that likely evolved under positive selection. 2D bar graphs are provided in the "Results-Current-Date/Results/Graphs" folder. These graphs display the positions of recurrently changing residues ("Posterior mean omega," top) and the probability of positive selection acting on each codon ("Prob. positive selection," middle, probability 0.7–1.0; "High prob. positive selection," bottom, probability 0.9–1.0). Codons excluded from the analysis are marked in gray. 3D representation of the most likely adaptive residues is found in the "Results-Current-Date/Results/Structures" folder, provided that the prediction model from the AlphaFold database matches the protein sequence extracted from the Ensembl database. For clarity, only residues with the highest probability (≥0.9) of having evolved under positive selection are mapped and their side chains are shown. The residues excluded from the analysis are colored in gray, and the N-terminal and C-terminal ends are labeled. Additionally, any domain annotation available in the Interpro database (Blum et al., 2021) is automatically marked with a distinct color and labeled allowing quick visual identification.

Manual alignment of structural prediction models

FREEDA-annotated protein structure prediction models from AlphaFold designated by their UniProt entries (MmMIS18β—A2AQ14; MmAURKC—O88445; MmINCENP—Q9WU62; MmCENP-O—Q8K015; MmCENP-P—Q9CZ92) were aligned to PDB entries (SpMIS18—5HJ0; HsAURKC—6GR8; HsINCENP—6GR8; HsCENP-

OP—7PB8) using an aligner module in PyMOL. Briefly, a FREEDA-annotated structure (e.g., "Aurkc_Mm.pse" found in "Results-Current-Date/Results/Structures") was opened in PyMOL, the selected PDB entry was downloaded (e.g., "fetch 6GR8"), and the two structures were aligned (e.g., "align Aurkc_Mm, 6GR8"). The align module first aligns protein sequences and then superimposes their structures, returning RMSD (Root-Mean-Square-Deviation). Lower RMSD values indicate better alignment. All alignments presented here returned RMSD below 2 Å.

Generation of CENP-O constructs

All CENP-O coding sequences were cloned from testis or liver samples. The use of rat CENP-O to represent a divergent ortholog was motivated by the availability of a rat (*Rattus norvegicus*) tissue sample for cloning. Chimeric CENP-O constructs were designed based on their 3D structure (AlphaFold database). Tissue was mechanically homogenized and total mRNA was isolated using TRIzol reagent (15596026; Invitrogen); cDNA was prepared using reverse transcription (18080051; SuperScript III First-Strand Synthesis System), amplified using construct-specific PCR primers (KK2502; KAPA HiFi Hot Start plus dNTPs; Roche), and inserted into the pGEMHE plasmid backbone (638948; In-fusion kit; Takara). Primers were designed using SnapGene (Dotmatics) software and are listed in Table S6. Each CENP-O construct was tagged with GFP at the C-terminus, separated by a linker of five glycines. Site-directed mutagenesis was performed using the Quik-Change Multisite Directed Mutagenesis kit (200515; Agilent) to introduce R225G, C226T, T228A, N247G, and P261H mutations. Rat CENP-O with 11 mouse point mutations was synthesized (Genewiz). The identity of all constructs was confirmed using Sanger sequencing of the entire coding sequence, including the reporter gene.

Oocyte isolation, microinjection, and in vitro maturation

Mus musculus mice (CF-1 strain) were purchased from Envigo NSA stock # 033. Females were primed with 5 U of pregnant mare somatic gonadotropin (367222; Calbiochem) injected into the intraperitoneal cavity 44–48 h prior to oocyte collections to induce superovulation. The ovaries were isolated using M2 medium (M7167; Sigma-Aldrich) supplemented with 2.5 mM of the maturation-blocking phosphodiesterase 3 inhibitor milrinone (M4659; 2.5 mM; Sigma-Aldrich Milipore). Germinal-vesicle oocytes were collected, denuded mechanically from cumulus cells, and incubated for at least 1 h prior to microinjection on a hot plate (38°C) under mineral oil (9305; FUJIFILM Irvine Scientific). Oocytes were then microinjected with ~5 pl of mRNAs in M2 medium with 2.5 mM milrinone and 3 mg/ml BSA at RT with a micromanipulator TransferMan NK 2 (Eppendorf) and a picoinjector (Medical Systems Corp). Oocytes were then incubated in 30–50 µl drops of Chatot-Ziomek-Bavister medium (MR019D; Thermo Fisher Scientific) under mineral oil (M5310; Sigma-Aldrich Milipore) at 37.8°C and 5% CO2 (Airgas) for 16 h to allow protein expression. The concentration of mRNA (15–30 ng/µl) used was selected to ensure similar cytoplasmic expression. mRNAs were synthesized using the T7 mScriptTM Standard mRNA Production System (C-MSC100625; CELLSCRIPT).

Immunofluorescence imaging

Oocyte maturation was induced in vitro by washing out milrinone 7.5 h before fixation. MI oocytes were fixed in freshly prepared 2% paraformaldehyde in PBS (pH 7.4) with 0.1% Triton X-100 for 20 min at RT, permeabilized in PBS with 0.1% Triton X-100 for 15 min at RT, placed in blocking solution (PBS with 0.3% BSA and 0.01% Tween-20) overnight at 4°C, incubated 1 h with primary antibody in blocking solution, washed three times for 15 min each, incubated 1 h with secondary antibody, washed three times for 15 min each, and mounted in Vectashield with DAPI (H-1200; Vector) to visualize chromosomes. Centromeres were labeled with CREST (human anti-human anti-centromere antibody, 1:200, HCT-0100; Immunovision) and an Alexa Fluor 594–conjugated goat anti-human secondary antibody (A-110014; Thermo Fisher Scientific). Confocal images were collected as 31 z-stacks at 0.5-µm intervals to visualize the entire meiotic spindle, using a confocal microscope (DMI4000B; Leica) equipped with a 63× 1.3 NA glycerol-immersion objective lens, an xy piezo Z stage (Applied Scientific Instrumentation), a spinning disk (Yokogawa Corporation of America), and an electron multiplier charge-coupled device camera (ImageEM C9100-13; Hamamatsu Photonics), controlled by MetaMorph software (Molecular Devices). Excitation was done with a Vortran Stradus VersaLase 4 laser module with 405-, 488-, 561-, and 639-nm lasers (Vortran Laser Technology). Panels of microscopic images were prepared using ImageJ (National Institutes of Health) free software.

Automated image analysis

Confocal images were analyzed using a custom-built Python-based automated program "Centrocalc" available here: https://github.com/DDudka9/Centrocalc. First, the program performs whole chromosome segmentation using the chromosome channel as a mask by grayscale dilation at a width of 16 pixels to ensure the centromeres are included. Then, a threshold is calculated using the Iterative Self-Organizing Data Analysis Technique method. If no chromosome channel is given, the entire cell is considered. Second, the program identifies centromeres using an approach based on (Vermolen et al., 2008). A difference of Gaussians algorithm is used to isolate spots of 150 nm. Third, a local maxima algorithm is used to identify centromeres. Up to 38 spots are chosen (the mouse 2n genome contains 40 chromosomes), separated by a minimum of two pixels by Chebyshev distance. The spots must be away from the edges of the image (20 pixels in x and y; 2 pixels in z). Fourth, 3D ellipsoid regions of interest (4 × 4 × 3 pixels) are drawn using the local maxima. Background regions of interest (ROIs) are drawn as volumes around the centromere ROIs (expanding the 3D ellipsoids by 1 pixel in all dimensions). Overlapping centromere ROIs are resolved by distance, where each pixel is assigned to the closest maxima. Fifth, centromere and background intensities are calculated as average grayscale pixel values and saved. ImageJ ROI files are also created to reference back to the original images. Modifying the Centrocalc source code can customize most of the described parameters.

Statistical analysis

Statistical analysis was performed using GraphPad Prism 9.3.1 (GraphPad Software). Statistical significance was assessed using a two-tailed *T* test (data distribution was normal under Kolmogorov–Smirnov and Shapiro–Wilk tests) or one-way ANOVA with Tukey's multiple comparison test (data distribution was assumed normal but was not formally tested). Graphs display means with standard deviations. P values indicated on graphs are $P \geq 0.05$, not significant (ns); $P < 0.05$, *; and $P < 0.0001$, ****.

Online supplemental material

Comparison of FREEDA with other pipelines is outlined in Fig. S1. Results of FREEDA validation are listed in Table S1 (validation of ortholog detection based on Ensembl database), Table S2 (accuracy in detecting positive selection using published datasets), Fig. S2, and Table S3 (accuracy in detecting sites under positive selection based on published data). Results of positive selection detection in rodent centromere and kinetochore proteins are shown in Table S4. The list of all genomic assemblies used by the FREEDA pipeline and the core packages comprising the pipeline are listed in Table S5. Table S6 lists all primers used to generate CENP-O constructs. Figs. S3, S4, S5, and S6 show structural analyses of selected centromeric proteins. Experimental analyses of rodent CENP-O are shown in Figs. S7, S9, and S10. Fig. S8 indicates sites that differ between mouse and rat CENP-O.

Data availability

FREEDA pipeline can be downloaded from the public GitHub repository https://github.com/DDudka9/freeda/releases. FREEDA documentation is available at https://ddudka9.github.io/freeda/. Supplementary materials including FREEDA validation results for all proteins analyzed in the manuscript are available on the open-access Zenodo repository at https://doi.org/10.5281/zenodo.7997737. These also include ortholog detection validation, analysis of rodent centromere proteins, and additional analyses of KIF4A, KIF4B, histone H4, MICA, MICB, NUP73, and HERC5 as mentioned in the text. Manually aligned structural prediction models for MIS18A, MIS18B, AURKC, CENP-O, and CENP-P are also included. Centrocalc is available here: https://github.com/DDudka9/Centrocalc.

Acknowledgments

We would like to thank all members of Michael Lampson's and Mia Levine's labs (University of Pennsylvania) for scientific discussions, especially Mia Levine and Cara Brand who provided critical feedback on the manuscript and FREEDA development. We extend special thanks to Piero Lamelza, Arunika Das, Hyuk-Joon Jeon, Cara Brand, Xueqi Sun, Sung-Ya Lin, Geng-Yuan Chen, and Katelyn Boese for testing the pipeline. Finally, we would like to acknowledge the vibrant community of the StackOverflow server for providing feedback on the code underlying our pipeline.

This work was supported by the Swiss National Science Foundation Early Postdoc Mobility grant P2GEP3_187772 awarded to D. Dudka, and National Institutes of Health grant R35GM122475 awarded to M.A. Lampson.

Author contributions: Conceptualization: D. Dudka and M.A. Lampson, Data curation: D. Dudka and R.B. Akins, Formal analysis: D. Dudka, Funding Acquisition: D. Dudka and M.A. Lampson, Investigation: D. Dudka and R.B. Akins, Project administration: D. Dudka and M.A. Lampson, Resources: M.A. Lampson, Software: D. Dudka and R.B. Akins, Supervision: D. Dudka and M.A. Lampson, Validation: D. Dudka and R.B. Akins, Visualization: D. Dudka and R.B. Akins, Methodology: D. Dudka and R.B. Akins, Writing–Original draft: D. Dudka and M.A. Lampson, Writing–Review and editing: D. Dudka, R.B. Akins and M.A. Lampson.

Disclosures: The authors declare no competing interests exist.

Submitted: 17 December 2022
Revised: 22 April 2023
Accepted: 7 June 2023

References

Abdul Azeez, K.R., S. Chatterjee, C. Yu, T.R. Golub, F. Sobott, and J.M. Elkins. 2019. Structural mechanism of synergistic activation of Aurora kinase B/C by phosphorylated INCENP. *Nat. Commun.* 10:3166. https://doi.org/10.1038/s41467-019-11085-0

Afanasyeva, A., M. Bockwoldt, C.R. Cooney, I. Heiland, and T.I. Gossmann. 2018. Human long intrinsically disordered protein regions are frequent targets of positive selection. *Genome Res.* 28:975–982. https://doi.org/10.1101/gr.232645.117

Akera, T., E. Trimm, and M.A. Lampson. 2019. Molecular strategies of meiotic cheating by selfish centromeres. *Cell.* 178:1132–1144.e10. https://doi.org/10.1016/j.cell.2019.07.001

Amaro, A.C., C.P. Samora, R. Holtackers, E. Wang, I.J. Kingston, M. Alonso, M. Lampson, A.D. McAinsh, and P. Meraldi. 2010. Molecular control of kinetochore-microtubule dynamics and chromosome oscillations. *Nat. Cell Biol.* 12:319–329. https://doi.org/10.1038/ncb2033

Anisimova, M., and D. Liberles. 2012. Detecting and understanding natural selection. In Codon Evolution: Mechanisms and Models. G.M. Cannarozzi, and A. Schneider, editors. Oxford University Press, Oxford. 73–96. https://doi.org/10.1093/acprof:osobl/9780199601165.003.0006

Anisimova, M., R. Nielsen, and Z. Yang. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics.* 164:1229–1236. https://doi.org/10.1093/genetics/164.3.1229

Balboula, A.Z., and K. Schindler. 2014. Selective disruption of aurora C kinase reveals distinct functions from aurora B kinase during meiosis in mouse oocytes. *PLoS Genet.* 10:e1004194. https://doi.org/10.1371/journal.pgen.1004194

Bancroft, J., P. Auckland, C.P. Samora, and A.D. McAinsh. 2015. Chromosome congression is promoted by CENP-Q- and CENP-E-dependent pathways. *J. Cell Sci.* 128:171–184.

Blum, M., H.Y. Chang, S. Chuguransky, T. Grego, S. Kandasaamy, A. Mitchell, G. Nuka, T. Paysan-Lafosse, M. Qureshi, S. Raj, et al. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49:D344–D354. https://doi.org/10.1093/nar/gkaa977

Brand, C.L., and M.T. Levine. 2021. Functional diversification of chromatin on rapid evolutionary timescales. *Annu. Rev. Genet.* 55:401–425. https://doi.org/10.1146/annurev-genet-071719-020301

Busset, J., C. Cabau, C. Meslin, and G. Pascal. 2011. PhyleasProg: A user-oriented web server for wide evolutionary analyses. *Nucleic Acids Res.* 39:W479–W485. https://doi.org/10.1093/nar/gkr243

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T.L. Madden. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics.* 10:421. https://doi.org/10.1186/1471-2105-10-421

Carlisle, J.A., and W.J. Swanson. 2020. Molecular mechanisms and evolution of fertilization proteins. *J. Exp. Zool. B Mol. Dev. Evol.* 336:652–665. https://doi.org/10.1002/jez.b.23004

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17:540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Chen, Q., M. Zhang, X. Pan, X. Yuan, L. Zhou, L. Yan, L.H. Zeng, J. Xu, B. Yang, L. Zhang, et al. 2021. Bub1 and CENP-U redundantly recruit Plk1 to stabilize kinetochore-microtubule attachments and ensure accurate chromosome segregation. *Cell Rep.* 36:109740. https://doi.org/10.1016/j.celrep.2021.109740

Cock, P.J., T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M.J. de Hoon. 2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 25:1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Cunningham, F., J.E. Allen, J. Allen, J. Alvarez-Jarreta, M.R. Amode, I.M. Armean, O. Austine-Orimoloye, A.G. Azov, I. Barnes, R. Bennett, et al. 2022. Ensembl 2022. *Nucleic Acids Res.* 50:D988–D995. https://doi.org/10.1093/nar/gkab1049

Dale, R.K., B.S. Pedersen, and A.R. Quinlan. 2011. Pybedtools: A flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*. 27:3423–3424. https://doi.org/10.1093/bioinformatics/btr539

Daugherty, M.D., and H.S. Malik. 2012. Rules of engagement: Molecular insights from host-virus arms races. *Annu. Rev. Genet.* 46:677–700. https://doi.org/10.1146/annurev-genet-110711-155522

Dudka, D., and M.A. Lampson. 2022. Centromere drive: Model systems and experimental progress. *Chromosome Res.* 30:187–203. https://doi.org/10.1007/s10577-022-09696-3

Edgar, R.C. 2004. Muscle: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797. https://doi.org/10.1093/nar/gkh340

Egan, A., A. Mahurkar, J. Crabtree, J.H. Badger, J.M. Carlton, and J.C. Silva. 2008. Idea: Interactive display for evolutionary analyses. *BMC Bioinformatics*. 9:524. https://doi.org/10.1186/1471-2105-9-524

Eskat, A., W. Deng, A. Hofmeister, S. Rudolphi, S. Emmerth, D. Hellwig, T. Ulbricht, V. Döring, J.M. Bancroft, A.D. McAinsh, et al. 2012. Step-wise assembly, maturation and dynamic behavior of the human CENP-P/O/R/Q/U kinetochore sub-complex. *PLoS One*. 7:e44717. https://doi.org/10.1371/journal.pone.0044717

Florio, M., M. Heide, A. Pinson, H. Brandl, M. Albert, S. Winkler, P. Wimberger, W.B. Huttner, and M. Hiller. 2018. Evolution and cell-type specificity of human-specific genes preferentially expressed in progenitors of fetal neocortex. *Elife*. 7:e32332. https://doi.org/10.7554/eLife.32332

Foltz, D.R., L.E. Jansen, B.E. Black, A.O. Bailey, J.R. Yates III, and D.W. Cleveland. 2006. The human CENP-A centromeric nucleosome-associated complex. *Nat. Cell Biol.* 8:458–469. https://doi.org/10.1038/ncb1397

Fujita, Y., T. Hayashi, T. Kiyomitsu, Y. Toyoda, A. Kokubu, C. Obuse, and M. Yanagida. 2007. Priming of centromere for CENP-A recruitment by human hMis18alpha, hMis18beta, and M18BP1. *Dev. Cell*. 12:17–30. https://doi.org/10.1016/j.devcel.2006.11.002

Gad, S.A., M. Sugiyama, M. Tsuge, K. Wakae, K. Fukano, M. Oshima, C. Sureau, N. Watanabe, T. Kato, A. Murayama, et al. 2022. The kinesin KIF4 mediates HBV/HDV entry through the regulation of surface NTCP localization and can be targeted by RXR agonists in vitro. *PLoS Pathog.* 18:e1009983. https://doi.org/10.1371/journal.ppat.1009983

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11:725–736. https://doi.org/10.1093/oxfordjournals.molbev.a040153

Gupta, R.K., S. Hué, T. Schaller, E. Verschoor, D. Pillay, and G.J. Towers. 2009. Mutation of a single residue renders human tetherin resistant to HIV-1 Vpu-mediated depletion. *PLoS Pathog.* 5:e1000443. https://doi.org/10.1371/journal.ppat.1000443

Hamming, R.W. 1950. Error detecting and error correcting codes. *Bell Syst. Tech. J.* 29:147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x

Henikoff, S., K. Ahmad, and H.S. Malik. 2001. The centromere paradox: Stable inheritance with rapidly evolving DNA. *Science*. 293:1098–1102. https://doi.org/10.1126/science.1062939

Hinshaw, S.M., and S.C. Harrison. 2019. The structure of the Ctf19c/CCAN from budding yeast. *Elife*. 8:e44239. https://doi.org/10.7554/eLife.44239

Hölzer, M., and M. Marz. 2021. PoSeiDon: A nextflow pipeline for the detection of evolutionary recombination events and positive selection. *Bioinformatics*. 37:1018–1020. https://doi.org/10.1093/bioinformatics/btaa695

Hongo, J.A., G.M. de Castro, L.C. Cintra, A. Zerlotini, and F.P. Lobo. 2015. Potion: An end-to-end pipeline for positive darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. *BMC Genomics*. 16:567. https://doi.org/10.1186/s12864-015-1765-0

Hori, T., M. Okada, K. Maenaka, and T. Fukagawa. 2008. CENP-O class proteins form a stable complex and are required for proper kinetochore function. *Mol. Biol. Cell*. 19:843–854. https://doi.org/10.1091/mbc.e07-06-0556

Jagadeeshan, S., and R.S. Singh. 2005. Rapidly evolving genes of Drosophila: Differing levels of selective pressure in testis, ovary, and head tissues between sibling species. *Mol. Biol. Evol.* 22:1793–1801. https://doi.org/10.1093/molbev/msi175

Jeffares, D.C., B. Tomiczek, V. Sojo, and M. dos Reis. 2015. A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. *Methods Mol. Biol.* 1201:65–90. https://doi.org/10.1007/978-1-4939-1438-8_4

Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*. 596:583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaessmann, H., N. Vinckenbosch, and M. Long. 2009. RNA-Based gene duplication: Mechanistic and evolutionary insights. *Nat. Rev. Genet.* 10:19–31. https://doi.org/10.1038/nrg2487

Kagawa, N., T. Hori, Y. Hoki, O. Hosoya, K. Tsutsui, Y. Saga, T. Sado, and T. Fukagawa. 2014. The CENP-O complex requirement varies among different cell types. *Chromosome Res.* 22:293–303. https://doi.org/10.1007/s10577-014-9404-1

Katoh, K., and D.M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780. https://doi.org/10.1093/molbev/mst010

Katoh, K., and D.M. Standley. 2016. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics*. 32:1933–1942. https://doi.org/10.1093/bioinformatics/btw108

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. Mafft: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Res.* 30:3059–3066. https://doi.org/10.1093/nar/gkf436

Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*. 267:275–276. https://doi.org/10.1038/267275a0

Kixmoeller, K., P.K. Allu, and B.E. Black. 2020. The centromere comes into focus: From CENP-A nucleosomes to kinetochore connections with the spindle. *Open Biol.* 10:200051. https://doi.org/10.1098/rsob.200051

Kops, G.J.P.L., and R. Gassmann. 2020. Crowning the kinetochore: The fibrous corona in chromosome segregation. *Trends Cell Biol.* 30:653–667. https://doi.org/10.1016/j.tcb.2020.04.006

Krenn, V., and A. Musacchio. 2015. The aurora B kinase in chromosome Bi-orientation and spindle checkpoint signaling. *Front. Oncol.* 5:225. https://doi.org/10.3389/fonc.2015.00225

Kumon, T., J. Ma, R.B. Akins, D. Stefanik, C.E. Nordgren, J. Kim, M.T. Levine, and M.A. Lampson. 2021. Parallel pathways for recruiting effector proteins determine centromere drive and suppression. *Cell*. 184:4904–4918.e11. https://doi.org/10.1016/j.cell.2021.07.037

Laguette, N., N. Rahm, B. Sobhian, C. Chable-Bessia, J. Münch, J. Snoeck, D. Sauter, W.M. Switzer, W. Heneine, F. Kirchhoff, et al. 2012. Evolutionary and functional analyses of the interaction between the myeloid restriction factor SAMHD1 and the lentiviral Vpx protein. *Cell Host Microbe*. 11:205–217. https://doi.org/10.1016/j.chom.2012.01.007

Lim, E.S., O.I. Fregoso, C.O. McCoy, F.A. Matsen, H.S. Malik, and M. Emerman. 2012. The ability of primate lentiviruses to degrade the monocyte restriction factor SAMHD1 preceded the birth of the viral accessory protein Vpx. *Cell Host Microbe*. 11:194–204. https://doi.org/10.1016/j.chom.2012.01.004

Liu, J., K. Chen, J.H. Wang, and C. Zhang. 2010. Molecular evolution of the primate antiviral restriction factor tetherin. *PLoS One*. 5:e11904. https://doi.org/10.1371/journal.pone.0011904

Löytynoja, A., and N. Goldman. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. USA*. 102:10557–10562. https://doi.org/10.1073/pnas.0409137102

Maeso, I., S.W. Roy, and M. Irimia. 2012. Widespread recurrent evolution of genomic features. *Genome Biol. Evol.* 4:486–500. https://doi.org/10.1093/gbe/evs022

Malik, H.S., D. Vermaak, and S. Henikoff. 2002. Recurrent evolution of DNA-binding motifs in the Drosophila centromeric histone. *Proc. Natl. Acad. Sci. USA*. 99:1449–1454. https://doi.org/10.1073/pnas.032664299

Mazumdar, M., S. Sundareshan, and T. Misteli. 2004. Human chromokinesin KIF4A functions in chromosome condensation and segregation. *J. Cell Biol.* 166:613–620. https://doi.org/10.1083/jcb.200401142

McKinley, K.L., and I.M. Cheeseman. 2016. The molecular basis for centromere identity and function. *Nat. Rev. Mol. Cell Biol.* 17:16–29. https://doi.org/10.1038/nrm.2015.5

Mellone, B.G., and D. Fachinetti. 2021. Diverse mechanisms of centromere specification. *Curr. Biol.* 31:R1491–R1504. https://doi.org/10.1016/j.cub.2021.09.083

Minoshima, Y., T. Hori, M. Okada, H. Kimura, T. Haraguchi, Y. Hiraoka, Y.C. Bao, T. Kawashima, T. Kitamura, and T. Fukagawa. 2005. The constitutive centromere component CENP-50 is required for recovery from spindle damage. *Mol. Cell. Biol.* 25:10315–10328. https://doi.org/10.1128/MCB.25.23.10315-10328.2005

Mitchell, P.S., C. Patzina, M. Emerman, O. Haller, H.S. Malik, and G. Kochs. 2012. Evolution-guided identification of antiviral specificity determinants in the broadly acting interferon-induced innate immunity factor MxA. *Cell Host Microbe*. 12:598–604. https://doi.org/10.1016/j.chom.2012.09.005

Muse, S.V., and B.S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11:715–724. https://doi.org/10.1093/oxfordjournals.molbev.a040152

Nguyen, L.H., T.T. Tran, L.T.N. Truong, H.H. Mai, and T.T. Nguyen. 2020. Overcharging of the zinc ion in the structure of the zinc-finger protein is needed for DNA binding stability. *Biochemistry*. 59:1378–1390. https://doi.org/10.1021/acs.biochem.9b01055

Nielsen, K., R.E. Marra, F. Hagen, T. Boekhout, T.G. Mitchell, G.M. Cox, and J. Heitman. 2005. Interaction between genetic background and the mating-type locus in Cryptococcus neoformans virulence potential. *Genetics*. 171:975–983. https://doi.org/10.1534/genetics.105.045039

Nilsson, J., M. Grahn, and A.P. Wright. 2011. Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. *Genome Biol.* 12:R65. https://doi.org/10.1186/gb-2011-12-7-r65

Okada, M., I.M. Cheeseman, T. Hori, K. Okawa, I.X. McLeod, J.R. Yates III, A. Desai, and T. Fukagawa. 2006. The CENP-H-I complex is required for the efficient incorporation of newly synthesized CENP-A into centromeres. *Nat. Cell Biol.* 8:446–457. https://doi.org/10.1038/ncb1396

Okonechnikov, K., O. Golosova, and M. Fursov, and UGENE team. 2012. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics*. 28:1166–1167. https://doi.org/10.1093/bioinformatics/bts091

Patel, M.R., Y.M. Loo, S.M. Horner, M. Gale Jr, and H.S. Malik. 2012. Convergent evolution of escape from hepaciviral antagonism in primates. *PLoS Biol.* 10:e1001282. https://doi.org/10.1371/journal.pbio.1001282

Peretti, D., L. Peris, S. Rosso, S. Quiroga, and A. Cáceres. 2000. Evidence for the involvement of KIF4 in the anterograde transport of L1-containing vesicles. *J. Cell Biol.* 149:141–152. https://doi.org/10.1083/jcb.149.1.141

Pesenti, M.E., D. Prumbaum, P. Auckland, C.M. Smith, A.C. Faesen, A. Petrovic, M. Erent, S. Maffini, S. Pentakota, J.R. Weir, et al. 2018. Reconstitution of a 26-subunit human kinetochore reveals cooperative microtubule binding by CENP-OPQUR and NDC80. *Mol. Cell.* 71:923–939.e10. https://doi.org/10.1016/j.molcel.2018.07.038

Pesenti, M.E., T. Raisch, D. Conti, K. Walstein, I. Hoffmann, D. Vogt, D. Prumbaum, I.R. Vetter, S. Raunser, and A. Musacchio. 2022. Structure of the human inner kinetochore CCAN complex and its significance for human centromere organization. *Mol. Cell.* 82:2113–2131.e8. https://doi.org/10.1016/j.molcel.2022.04.027

Picard, L., Q. Ganivet, O. Allatif, A. Cimarelli, L. Guéguen, and L. Etienne. 2020. DGINN, an automated and highly-flexible pipeline for the detection of genetic innovations on protein-coding genes. *Nucleic Acids Res.* 48:e103. https://doi.org/10.1093/nar/gkaa680

Pond, S.L., S.D. Frost, and S.V. Muse. 2005. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*. 21:676–679. https://doi.org/10.1093/bioinformatics/bti079

Quan, S., I. Schneider, J. Pan, A. Von Hacht, and J.C.A. Bardwell. 2007. The CXXC motif is more than a redox rheostat. *J. Biol. Chem.* 282:28823–28833. https://doi.org/10.1074/jbc.M705291200

Ranwez, V., S. Harispe, F. Delsuc, and E.J. Douzery. 2011. Macse: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS One*. 6:e22594. https://doi.org/10.1371/journal.pone.0022594

Ridout, K.E., C.J. Dixon, and D.A. Filatov. 2010. Positive selection differs between protein secondary structure elements in Drosophila. *Genome Biol. Evol.* 2:166–179. https://doi.org/10.1093/gbe/evq008

Rosin, L., and B.G. Mellone. 2016. Co-Evolving CENP-A and CAL1 domains mediate centromeric CENP-A deposition across Drosophila species. *Dev. Cell.* 37:136–147. https://doi.org/10.1016/j.devcel.2016.03.021

Roxström-Lindquist, K., and I. Faye. 2001. The Drosophila gene Yippee reveals a novel family of putative zinc binding proteins highly conserved among eukaryotes. *Insect Mol. Biol.* 10:77–86. https://doi.org/10.1046/j.1365-2583.2001.00239.x

Sahm, A., M. Bens, M. Platzer, and K. Szafranski. 2017. PosiGene: Automated and easy-to-use pipeline for genome-wide detection of positively selected genes. *Nucleic Acids Res.* 45:e100. https://doi.org/10.1093/nar/gkx179

Sawyer, S.L., L.I. Wu, M. Emerman, and H.S. Malik. 2005. Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proc. Natl. Acad. Sci. USA*. 102:2832–2837. https://doi.org/10.1073/pnas.0409853102

Sayers, E.W., J. Beck, E.E. Bolton, D. Bourexis, J.R. Brister, K. Canese, D.C. Comeau, K. Funk, S. Kim, W. Klimke, et al. 2021. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 49:D10–D17. https://doi.org/10.1093/nar/gkaa892

Schmitzberger, F., and S.C. Harrison. 2012. RWD domain: A recurring module in kinetochore architecture shown by a ctf19-Mcm21 complex structure. *EMBO Rep.* 13:216–222. https://doi.org/10.1038/embor.2012.1

Sievers, F., A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539. https://doi.org/10.1038/msb.2011.75

Singh, P., M.E. Pesenti, S. Maffini, S. Carmignani, M. Hedtfeld, A. Petrovic, A. Srinivasamani, T. Bange, and A. Musacchio. 2021. BUB1 and CENP-U, primed by CDK1, are the main PLK1 kinetochore receptors in mitosis. *Mol. Cell.* 81:67–87.e9. https://doi.org/10.1016/j.molcel.2020.10.040

Sironi, M., R. Cagliani, D. Forni, and M. Clerici. 2015. Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* 16:224–236. https://doi.org/10.1038/nrg3905

Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stapley, J., P.G.D. Feulner, S.E. Johnston, A.W. Santure, and C.M. Smadja. 2017. Variation in recombination frequency and distribution across eukaryotes: Patterns and processes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 372:20160455. https://doi.org/10.1098/rstb.2016.0455

Starr, T.N., and J.W. Thornton. 2016. Epistasis in protein evolution. *Protein Sci.* 25:1204–1218. https://doi.org/10.1002/pro.2897

Steinway, S.N., R. Dannenfelser, C.D. Laucius, J.E. Hayes, and S. Nayak. 2010. JCoDA: A tool for detecting evolutionary selection. *BMC Bioinformatics*. 11:284. https://doi.org/10.1186/1471-2105-11-284

Stellfox, M.E., I.K. Nardi, C.M. Knippler, and D.R. Foltz. 2016. Differential binding partners of the Mis18α/β YIPPEE domains regulate Mis18 complex recruitment to centromeres. *Cell Rep.* 15:2127–2135. https://doi.org/10.1016/j.celrep.2016.05.004

Stern, A., A. Doron-Faigenboim, E. Erez, E. Martz, E. Bacharach, and T. Pupko. 2007. Selecton 2007: Advanced models for detecting positive and purifying selection using a bayesian inference approach. *Nucleic Acids Res.* 35:W506–W511. https://doi.org/10.1093/nar/gkm382

Stremlau, M., M. Perron, S. Welikala, and J. Sodroski. 2005. Species-specific variation in the B30.2(SPRY) domain of TRIM5alpha determines the potency of human immunodeficiency virus restriction. *J. Virol.* 79:3139–3145. https://doi.org/10.1128/JVI.79.5.3139-3145.2005

Subramanian, L., B. Medina-Pritchard, R. Barton, F. Spiller, R. Kulasegaran-Shylini, G. Radaviciute, R.C. Allshire, and A. Arockia Jeyaprakash. 2016. Centromere localization and function of Mis18 requires Yippee-like domain-mediated oligomerization. *EMBO Rep.* 17:496–507. https://doi.org/10.15252/embr.201541520

Swanson, W.J., and V.D. Vacquier. 2002. The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* 3:137–144. https://doi.org/10.1038/nrg733

Talavera, G., and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* 56:564–577. https://doi.org/10.1080/10635150701472164

Tamura, K., G. Stecher, and S. Kumar. 2021. MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* 38:3022–3027. https://doi.org/10.1093/molbev/msab120

Taylor, S.S., and A.P. Kornev. 2011. Protein kinases: Evolution of dynamic regulatory proteins. *Trends Biochem. Sci.* 36:65–77. https://doi.org/10.1016/j.tibs.2010.09.006

Tromer, E.C., J.J.E. van Hooff, G.J.P.L. Kops, and B. Snel. 2019. Mosaic origin of the eukaryotic kinetochore. *Proc. Natl. Acad. Sci. USA*. 116:12873–12882. https://doi.org/10.1073/pnas.1821945116

UniProt Consortium. 2021. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49:D480–D489. https://doi.org/10.1093/nar/gkaa1100

Ustianenko, D., S.M. Weyn-Vanhentenryck, and C. Zhang. 2017. Microexons: Discovery, regulation, and function. *Wiley Interdiscip. Rev. RNA*. 8. https://doi.org/10.1002/wrna.1418

van der Lee, R., L. Wiel, T.J.P. van Dam, and M.A. Huynen. 2017. Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. *Nucleic Acids Res.* 45:10634–10648. https://doi.org/10.1093/nar/gkx704

Vermaak, D., H.S. Hayden, and S. Henikoff. 2002. Centromere targeting element within the histone fold domain of Cid. *Mol. Cell. Biol.* 22:7553–7561. https://doi.org/10.1128/MCB.22.21.7553-7561.2002

Vermolen, B.J., Y. Garini, I.T. Young, R.W. Dirks, and V. Raz. 2008. Segmentation and analysis of the three-dimensional redistribution of nuclear components in human mesenchymal stem cells. *Cytometry A*. 73: 816–824. https://doi.org/10.1002/cyto.a.20612

Wilfert, L., J. Gadau, and P. Schmid-Hempel. 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity*. 98:189–197. https://doi.org/10.1038/sj.hdy.6800950

Wu, T., S.I.R. Lane, S.L. Morgan, and K.T. Jones. 2018. Spindle tubulin and MTOC asymmetries may explain meiotic drive in oocytes. *Nat. Commun.* 9:2952. https://doi.org/10.1038/s41467-018-05338-7

Yang, Z. 2007. Paml 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586–1591. https://doi.org/10.1093/molbev/msm088

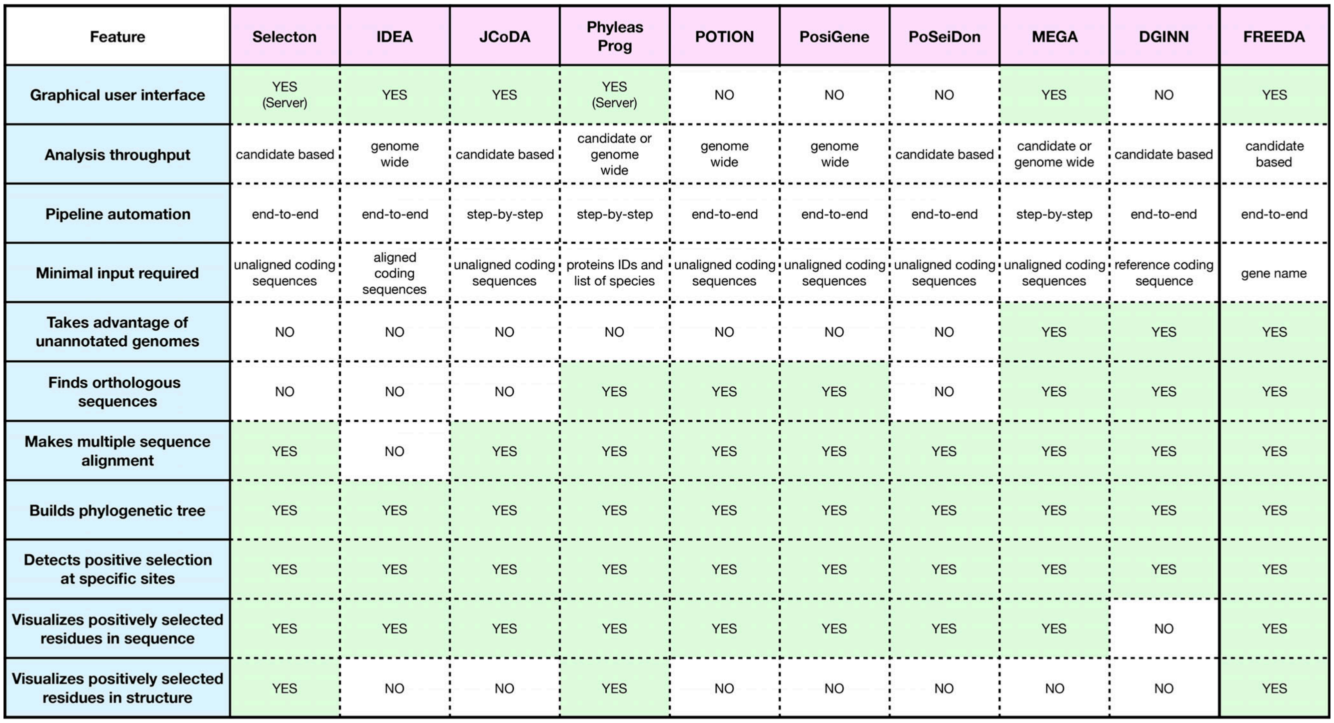
Yap, M.W., S. Nisole, and J.P. Stoye. 2005. A single amino acid change in the SPRY domain of human Trim5alpha leads to HIV-1 restriction. *Curr. Biol.* 15:73–78. https://doi.org/10.1016/j.cub.2004.12.042

Yatskevich, S., K.W. Muir, D. Bellini, Z. Zhang, J. Yang, T. Tischer, M. Predin, T. Dendooven, S.H. McLaughlin, and D. Barford. 2022. Structure of the human inner kinetochore bound to a centromeric CENP-A nucleosome. *Science*. 376:844–852. https://doi.org/10.1126/science.abn3810

Zasadzińska, E., and D.R. Foltz. 2017. Orchestrating the specific assembly of centromeric nucleosomes. *Prog. Mol. Subcell. Biol.* 56:165–192. https://doi.org/10.1007/978-3-319-58592-5_7

Zhou, W., D. Richmond-Buccola, Q. Wang, and P.J. Kranzusch. 2022. Structural basis of human TREX1 DNA degradation and autoimmune disease. *Nat. Commun.* 13:4277. https://doi.org/10.1038/s41467-022-32055-z

Supplemental material



Feature	Selecton	IDEA	JCoDA	Phyleas Prog	POTION	PosiGene	PoSeiDon	MEGA	DGINN	FREEDA
Graphical user interface	YES (Server)	YES	YES	YES (Server)	NO	NO	NO	YES	NO	YES
Analysis throughput	candidate based	genome wide	candidate based	candidate or genome wide	genome wide	genome wide	candidate based	candidate or genome wide	candidate based	candidate based
Pipeline automation	end-to-end	end-to-end	step-by-step	step-by-step	end-to-end	end-to-end	end-to-end	step-by-step	end-to-end	end-to-end
Minimal input required	unaligned coding sequences	aligned coding sequences	unaligned coding sequences	proteins IDs and list of species	unaligned coding sequences	unaligned coding sequences	unaligned coding sequences	unaligned coding sequences	reference coding sequence	gene name
Takes advantage of unannotated genomes	NO	NO	NO	NO	NO	NO	NO	YES	YES	YES
Finds orthologous sequences	NO	NO	NO	YES	YES	YES	NO	YES	YES	YES
Makes multiple sequence alignment	YES	NO	YES	YES	YES	YES	YES	YES	YES	YES
Builds phylogenetic tree	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Detects positive selection at specific sites	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Visualizes positively selected residues in sequence	YES	YES	YES	YES	YES	YES	YES	YES	NO	YES
Visualizes positively selected residues in structure	YES	NO	NO	YES	NO	NO	NO	NO	NO	YES

Figure S1. **Features of published automated pipelines used to detect signatures of positive selection.** Only automated pipelines are represented: Selecton (Stern et al., 2007), IDEA (Egan et al., 2008), JCoDA (Steinway et al., 2010), PhyleasProg (Busset et al., 2011), POTION (Hongo et al., 2015), PosiGene (Sahm et al., 2017), PoSeiDon (Hölzer and Marz, 2021), MEGA (Tamura et al., 2021), and DGINN (Picard et al., 2020).

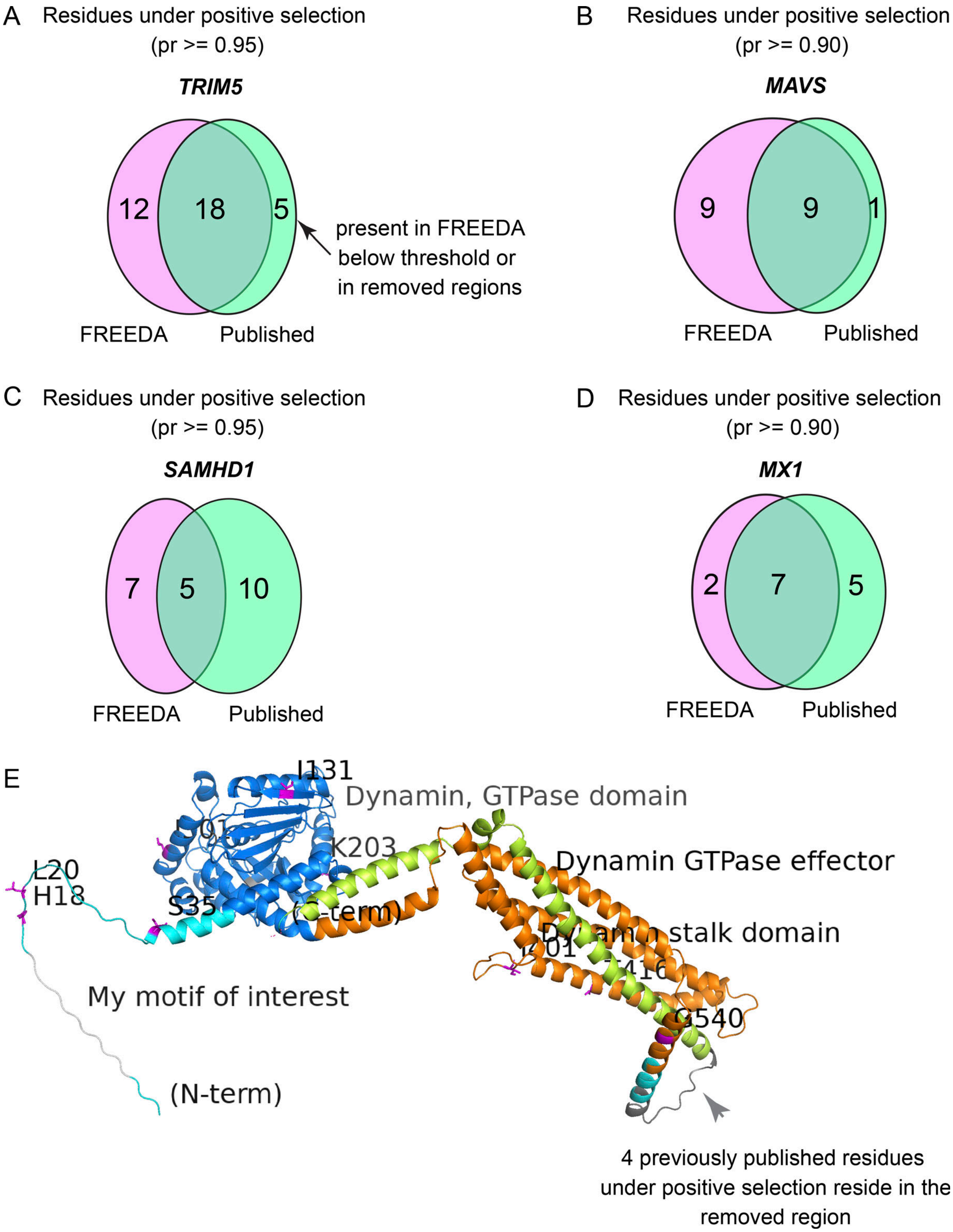


Figure S2. **Pipeline validation at the level of single residues. (A)** *TRIM5* analysis of 16 species with 96% coding sequence (CDS) coverage as compared with Sawyer et al. (2005) and van der Lee et al. (2017). **(B)** *MAVS* analysis of 19 species with 97% CDS coverage as compared with Patel et al. (2012). **(C)** *SAMHD1* analysis of 18 species with 92% CDS coverage as compared with Laguette et al. (2012), Lim et al. (2012), and van der Lee et al. (2017). **(D)** *MX1* analysis of 18 species with 97% CDS coverage as compared with Mitchell et al. (2012). See Table S3 for detailed analyses. Only the number of residues from M8 vs. M7 analysis are reported (see Materials and methods for details). Comparisons are made using similar phylogenetic branches of *Simiiformes* (hominoids, Old World monkeys, and New World monkeys) in most cases. Probability thresholds are set to match those used by the referenced studies. Number of species used varies due to the elimination of incomplete sequences from FREEDA analyses. Magenta ellipses, number of sites matching the threshold found by FREEDA; green ellipses, number of published sites. **(E)** Raw FREEDA-annotated image of the MxA protein. Note the region removed from the analysis due to alignment uncertainty (dark grey arrowhead) where four residues under positive selection have previously been identified (Mitchell et al., 2012).

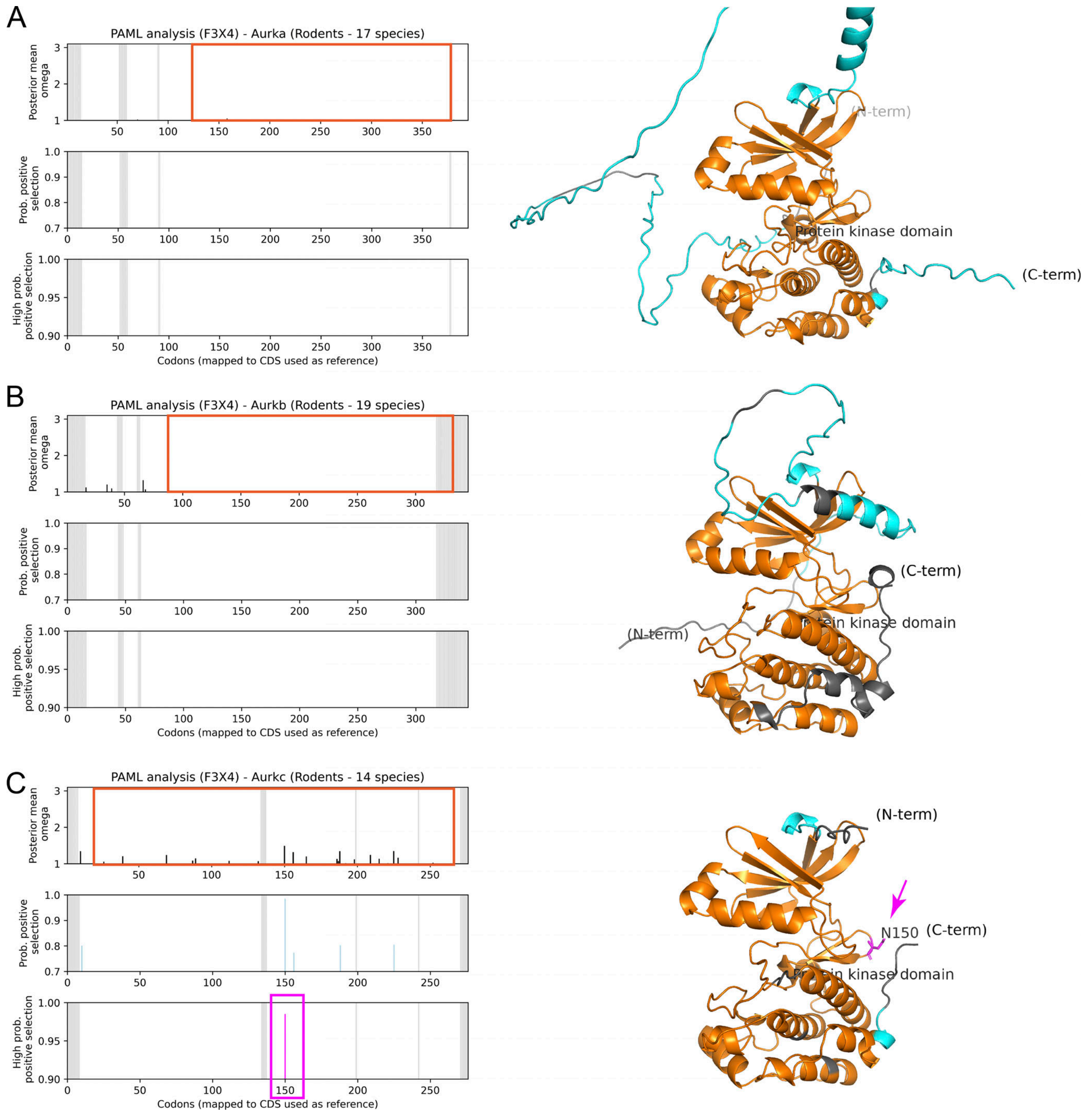


Figure S3. **Aurora kinases differ in the number of recurrently changing residues. (A–C)** Recurrently changing residues, mapped onto reference coding sequences, are not detected in AURKA (A) and AURKB (B), as compared with AURKC (C). Orange frames: protein kinase domain. Structural models highlight high structural conservation of Aurora protein kinase domains (orange). Parts of the AURKA N-terminus were cropped to allow better visualization and comparison of kinase domains.

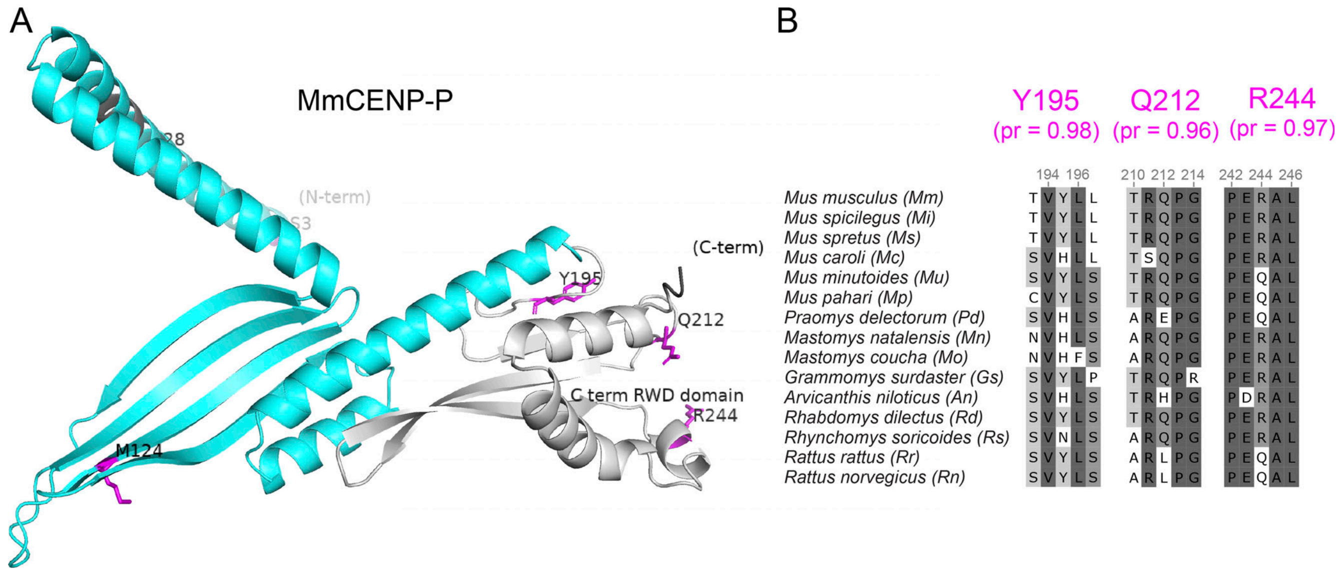

A
MmCENP-P
88
(N-term)
S3
(C-term)
Y195
Q212
C term RWD domain
R244
M124
B
Y195
(pr = 0.98)
Q212
(pr = 0.96)
R244
(pr = 0.97)
194 196 210 212 214 242 244 246
Mus musculus (Mm) T V Y L L T R Q P G P E R A L
Mus spicilegus (Mi) T V Y L L T R Q P G P E R A L
Mus spretus (Ms) T V Y L L T R Q P G P E R A L
Mus caroli (Mc) S V H L L T S Q P G P E R A L
Mus minutoides (Mu) S V Y L S T R Q P G P E Q A L
Mus pahari (Mp) C V Y L S T R Q P G P E Q A L
Praomys delectorum (Pd) S V H L S A R E P G P E Q A L
Mastomys natalensis (Mn) N V H L S A R Q P G P E R A L
Mastomys coucha (Mo) N V H F S A R Q P G P E R A L
Grammomys surdaster (Gs) S V Y L P T R Q P R P E R A L
Arvicanthis niloticus (An) S V H L S T R H P G P D R A L
Rhabdomys dilectus (Rd) S V Y L S T R Q P G P E R A L
Rhynchomys soricoides (Rs) S V N L S A R Q P G P E R A L
Rattus rattus (Rr) S V Y L S A R L P G P E Q A L
Rattus norvegicus (Rn) S V Y L S A R L P G P E Q A L

Figure S4. **Signatures of positive selection in the CENP-P C-terminal RWD domain. (A)** Annotated structural prediction model of MmCENP-P, generated automatically by FREEDA and visualized in PyMOL without manual modifications. The C-terminal RWD domain (labeled by the user within the GUI) is in gray. **(B)** Three residues (magenta) with the highest probability of having evolved under positive selection (≥0.9) are shown in snippets of the multiple sequence alignment in *Murinae*. Dark gray: highly conserved residues, gray: less conserved residues, white: non-synonymous substitutions.

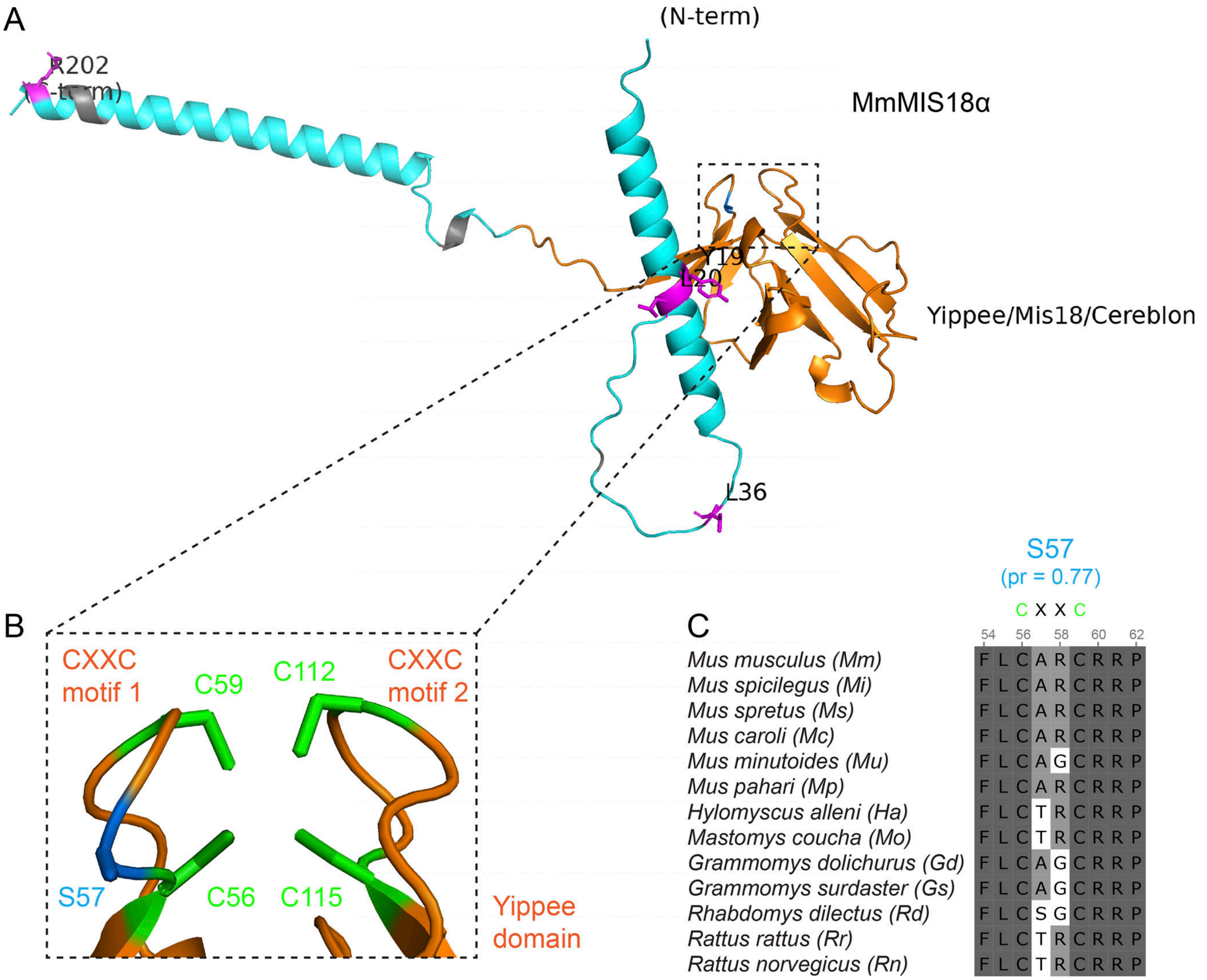

A
(N-term)
R202
(C-term)
MmMIS18α
Y19
L20
Yippee/Mis18/Cereblon
L36
B
CXXC
motif 1
C59
C112
CXXC
motif 2
S57
C56
C115
Yippee
domain
C
S57
(pr = 0.77)
C X X C
54 56 58 60 62
Mus musculus (Mm) F L C A R C R R P
Mus spicilegus (Mi) F L C A R C R R P
Mus spretus (Ms) F L C A R C R R P
Mus caroli (Mc) F L C A R C R R P
Mus minutoides (Mu) F L C A G C R R P
Mus pahari (Mp) F L C A R C R R P
Hylomyscus alleni (Ha) F L C T R C R R P
Mastomys coucha (Mo) F L C T R C R R P
Grammomys dolichurus (Gd) F L C A G C R R P
Grammomys surdaster (Gs) F L C A G C R R P
Rhabdomys dilectus (Rd) F L C S G C R R P
Rattus rattus (Rr) F L C T R C R R P
Rattus norvegicus (Rn) F L C T R C R R P

Figure S5. **Signatures of positive selection in the CXXC motif of the Yippee domain of MIS18α. (A)** Annotated structural prediction model of mouse MIS18α (MmMIS18α), generated automatically by FREEDA and visualized in PyMOL without any manual modifications. Orange: Yippee domain, magenta: highly likely adaptive residues (probability ≥ 0.9). **(B)** Enlarged CXXC motifs with the S57 residue (blue) within motif 1. Green: conserved cysteine residues. **(C)** Snippet of the multiple sequence alignment of CXXC motif 1 in *Murinae*. Dark gray: highly conserved residues, gray: less conserved residues, white: non-synonymous substitutions.

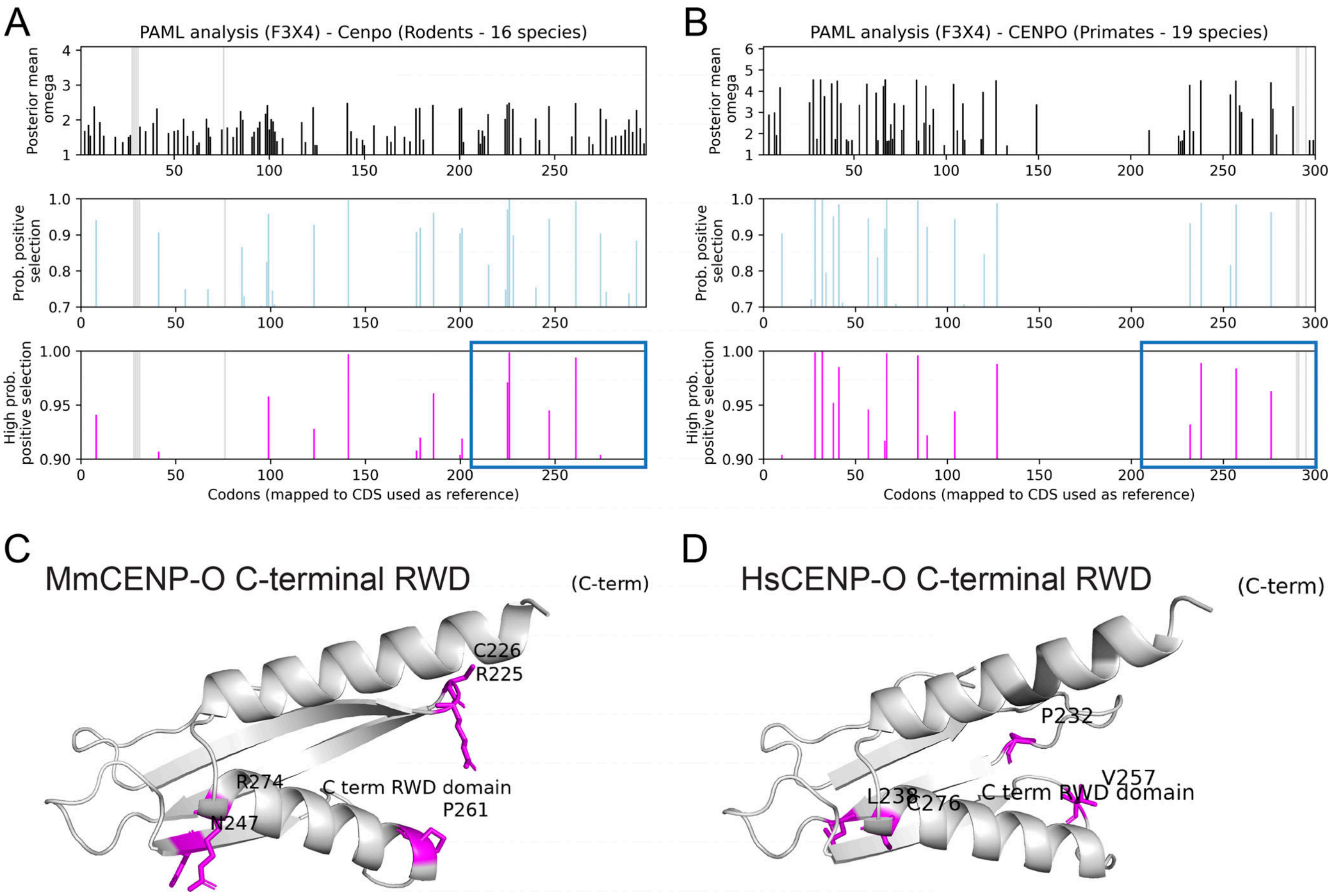


Figure S6. **Signatures of positive selection in loops and turns of rodent and primate CENP-O C-terminal RWD domains. (A and B)** Most likely adaptive residues in rodent (A) or primate (B) CENP-O mapped to the mouse or human CENP-O coding sequence, respectively, by FREEDA. Blue frames were manually added to outline C-terminal RWD domains. **(C and D)** FREEDA-annotated C-terminal RWD domains of mouse (C) and human (D) CENP-O. Magenta: highly likely adaptive residues (probability ≥ 0.9).

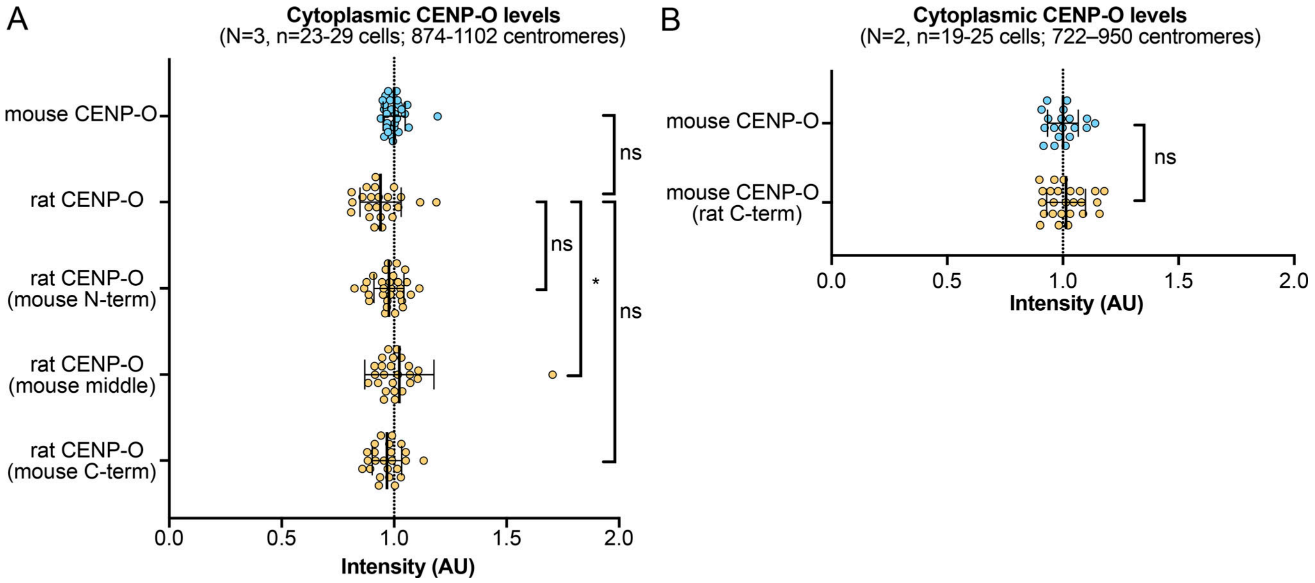


Figure S7. **Expression levels of microinjected CENP-O–GFP are equal between constructs. (A and B)** Quantification of CENP-O–GFP levels in cytoplasm for constructs analyzed in Fig. 6, A and B (A) and Fig. 6, C and D (B). Each spot represents one cell. For each construct, 38 cytoplasmic ROIs per cell were analyzed (see Materials and methods for details) from ≥19 cells from three (A) or two (B) independent experiments. Bars: mean intensity with standard deviation. *P < 0.05; ns: not significant; one-way ANOVA with Tukey's multiple comparison test (A) or two-tailed Student's *T* test (B).

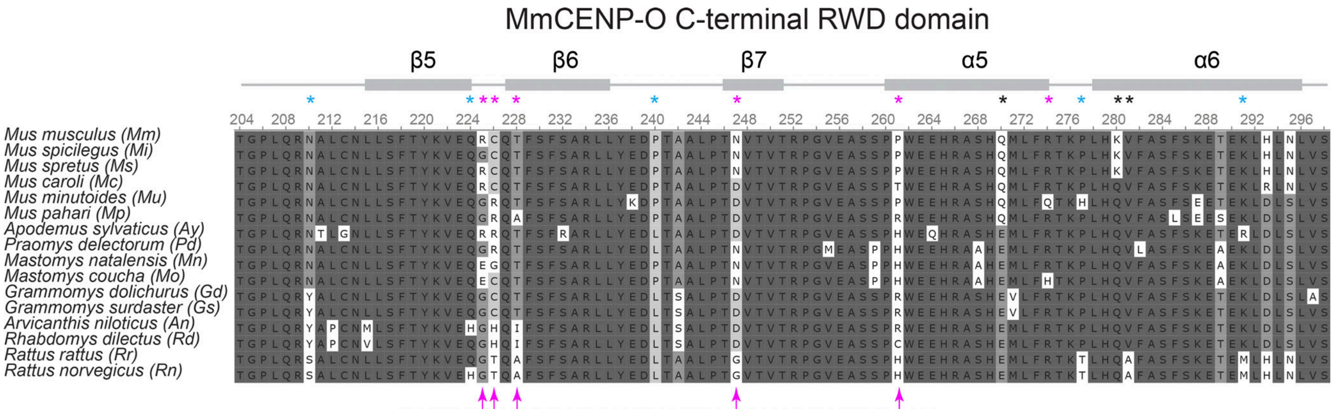

MmCENP-O C-terminal RWD domain
β5
β6
β7
α5
α6
204 208 212 216 220 224 228 232 236 240 244 248 252 256 260 264 268 272 276 280 284 288 292 296
Mus musculus (Mm) TGPLQRNALCNLLSFTYKVEQRCQTFSFSARLLYEDPTAALPTNVTVTRPGVEASSPPWEEHRASHQMLFRTKPLHKVFASFSKETEKLHLNLVS
Mus spicilegus (Mi) TGPLQRNALCNLLSFTYKVEQGCQTFSFSARLLYEDPTAALPTNVTVTRPGVEASSPPWEEHRASHQMLFRTKPLHKVFASFSKETEKLHLNLVS
Mus spretus (Ms) TGPLQRNALCNLLSFTYKVEQRCQTFSFSARLLYEDPTAALPTNVTVTRPGVEASSPPWEEHRASHQMLFRTKPLHKVFASFSKETEKLHLNLVS
Mus caroli (Mc) TGPLQRNALCNLLSFTYKVEQRCQTFSFSARLLYEDPTAALPTDVTVTRPGVEASSPTWEEHRASHQMLFRTKPLHQVFASFSKETEKLRLNLVS
Mus minutoides (Mu) TGPLQRNALCNLLSFTYKVEQGRQTFSFSARLLYKDPTAALPTDVTVTRPGVEASSPPWEEHRASHQMLFQTKHLHQVFASFSEETEKLDLSLVS
Mus pahari (Mp) TGPLQRNALCNLLSFTYKVEQGRQAFSFSARLLYEDPTAALPTDVTVTRPGVEASSPRWEEHRASHQMLFRTKPLHQVFASLSEESEKLDLSLVS
Apodemus sylvaticus (Ay) TGPLQRNTLGNLLSFTYKVEQRRQTFSFRARLLYEDLTAALPTDVTVTRPGVEASSPHWEQHRASHEMLFRTKPLHQVFASFSKETERLDLSLVS
Praomys delectorum (Pd) TGPLQRNALCNLLSFTYKVEQGRQTFSFSARLLYEDLTAALPTNVTVTRPGMEASPPHWEEHRAAHEMLFRTKPLHQVLASFSKEAEKLDLSLVS
Mastomys natalensis (Mn) TGPLQRNALCNLLSFTYKVEQEGQTFSFSARLLYEDPTAALPTNVTVTRPGVEASPPHWEEHRAAHEMLFRTKPLHQVFASFSKEAEKLDLSLVS
Mastomys coucha (Mo) TGPLQRNALCNLLSFTYKVEQECQTFSFSARLLYEDPTAALPTNVTVTRPGVEASPPHWEEHRAAHEMLFHTKPLHQVFASFSKEAEKLDLSLVS
Grammomys dolichurus (Gd) TGPLQRYALCNLLSFTYKVEQGCQTFSFSARLLYEDLTSALPTDVTVTRPGVEASSPRWEEHRASHEVLFRTKPLHQVFASFSKETEKLDLSLAS
Grammomys surdaster (Gs) TGPLQRYALCNLLSFTYKVEQGCQTFSFSARLLYEDLTSALPTDVTVTRPGVEASSPRWEEHRASHEVLFRTKPLHQVFASFSKETEKLDLSLVS
Arvicanthis niloticus (An) TGPLQRYAPCNMLSFTYKVEHGHQIFSFSARLLYEDLTSALPTDVTVTRPGVEASSPRWEEHRASHEMLFRTKPLHQVFASFSKETEKLDLSLVS
Rhabdomys dilectus (Rd) TGPLQRYAPCNVLSFTYKVEQGHQIFSFSARLLYEDLTSALPTDVTVTRPGVEASSPCWEEHRASHEMLFRTKPLHQVFASFSKETEKLDLSLVS
Rattus rattus (Rr) TGPLQRSALCNLLSFTYKVEQGTQAFSFSARLLYEDLTAALPTGVTVTRPGVEASSPHWEEHRASHEMLFRTKTLHQAFASFSKETEMLHLNLVS
Rattus norvegicus (Rn) TGPLQRSALCNLLSFTYKVEHGTQAFSFSARLLYEDLTAALPTGVTVTRPGVEASSPHWEEHRASHEMLFRTKTLHQAFASFSKETEMLHLNLVS

Figure S8. **Multiple sequence alignment of the rodent CENP-O C-terminal domain.** Schematic shows α-helices and β-sheets (gray boxes) and loops and turns (gray lines). Asterisks indicate highly likely adaptive residues (magenta, probability ≥ 0.9), less likely adaptive residues (blue, probability ≥ 0.5), and residues that differ between mouse and rat but do not evolve under positive selection (black). Magenta arrows indicate residues swapped in the experiment. Alignment: dark gray: highly conserved residues, gray: less conserved residues, white: non-synonymous substitutions.

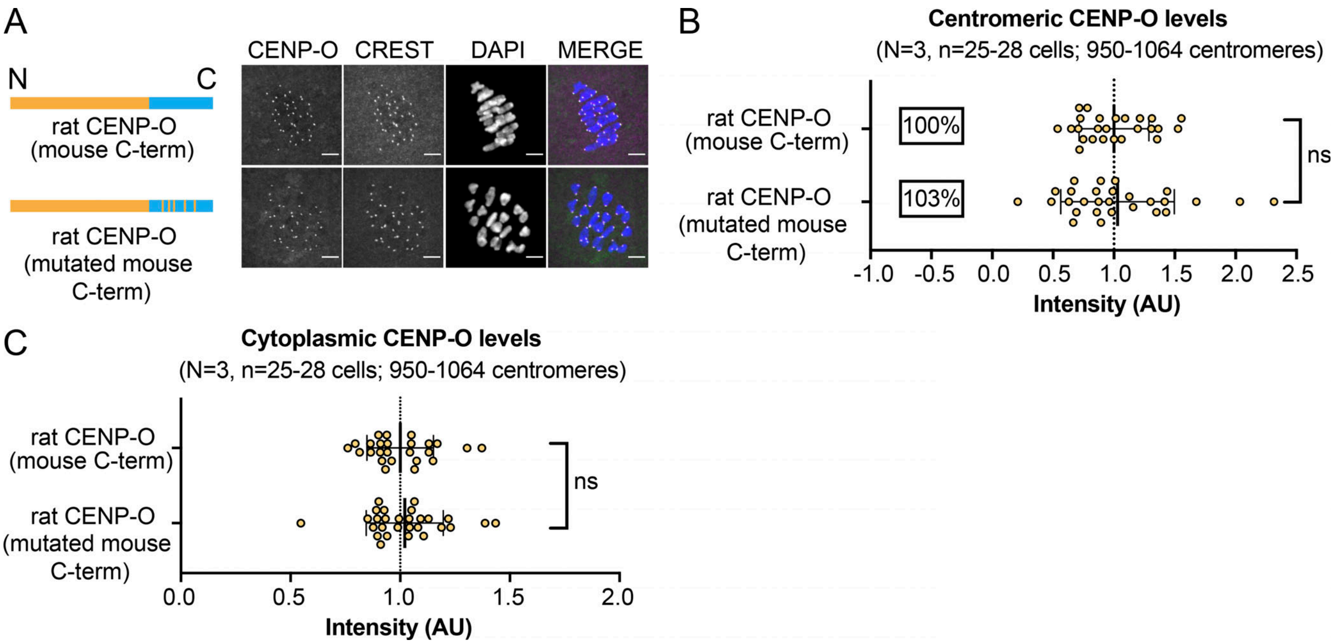


Figure S9. **Swapping five most likely adaptive residues in the CENP-O C-terminal RWD domain is insufficient to reduce centromere binding.** Mouse oocytes expressing the indicated CENP-O–GFP constructs were fixed in meiosis I and stained for centromeres (CREST) and DNA (DAPI). One construct is rat CENP-O with the mouse C-terminal RWD domain, and the other is identical except for five mutations in the RWD domain swapping mouse-specific residues with the highest probability of having evolved under positive selection (probability ≥ 0.9; R225G, C226T, T228A, N247G, P261H) to the corresponding rat-specific residues. **(A–C)** Images (A) show maximum projections; scale bars, 5 µm. Graphs show CENP-O-GFP intensity at centromeres (B) or in the cytoplasm (C); for each construct, *n* ≥ 950 centromeres from ≥25 cells from three independent experiments. Each spot represents one cell; bars, mean intensities with standard deviation; ns: not significant, two-tailed Student's *T* test.

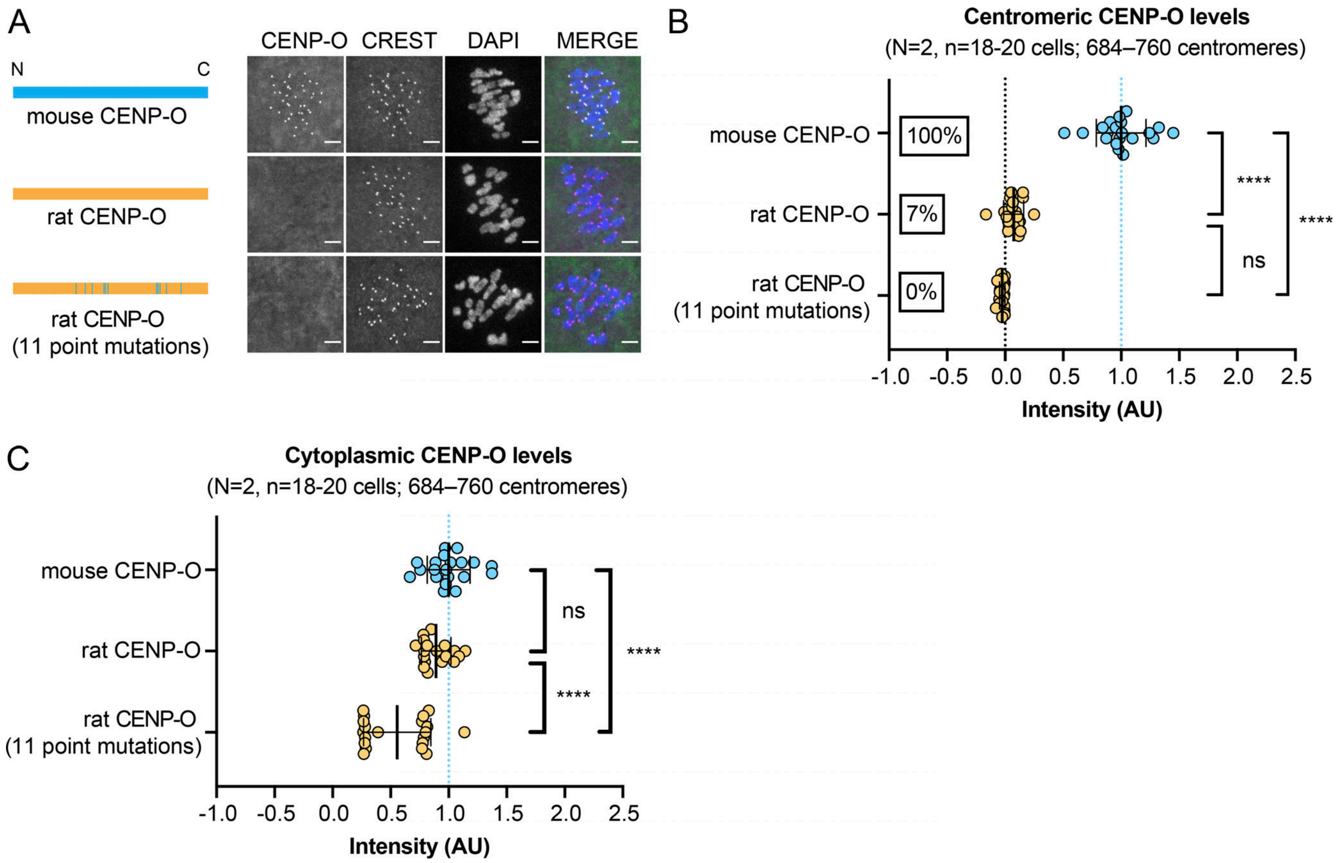


Figure S10. **Swapping 11 most likely adaptive residues in rat CENP-O is insufficient to rescue centromere binding.** Mouse oocytes expressing the indicated CENP-O–GFP constructs were fixed in meiosis I and stained for centromeres (CREST) and DNA (DAPI). The constructs are: mouse CENP-O, rat CENP-O, and rat CENP-O with 11 mouse-specific residues with the highest probability of having evolved under positive selection (probability ≥ 0.9; L99V, Y123C, A141V, S177R, S179W, V186A, G225R, T226C, A228T, G247N, and H261P). **(A–C)** Images (A) show maximum projections; scale bars, 5 μm. Graphs show CENP-O–GFP intensity at centromeres (B) or in the cytoplasm (C); for each construct, *n* ≥ 684 centromeres from ≥18 cells from two independent experiments. Each spot represents one cell; bars, mean intensities with standard deviation; ****P < 0.0001, ns: not significant, one-way ANOVA with Tukey's multiple comparison test.

Provided online are six tables. Table S1 shows testing FREEDA accuracy in detecting orthologs. Table S2 shows FREEDA results analyzing test datasets comprising 23 primate proteins. Table S3 shows the comparison of specific sites that likely evolved under positive selection published previously and those detected by FREEDA in selected genes. Table S4 shows FREEDA results analyzing 104 centromeric proteins in rodents. Table S5 lists genomic assemblies and core packages used by FREEDA to detect statistical signatures of positive selection. Table S6 lists primers used for generating CENP-O constructs used in this study.