Title: Naturally arising *de novo* open reading frames as potential zinc chelators in *Drosophila melanogaster*

Authors: UnJin Lee[1]*

[1]Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY, 10065, USA

*Correspondence to: ulee@rockefeller.edu

Classification: Biological Sciences — Evolution (primary); Genetics (secondary).

Keywords: de novo gene origination; microsatellite emergence bias and evolvability; bis-histidine metal chelation; zinc homeostasis in Drosophila; METTL9 post-translational regulation.

Author contributions: U.L. designed research, performed research, analyzed data, and wrote the paper.

Corresponding author: UnJin Lee, Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, 1230 York Avenue, New York, NY 10065, USA; email: ulee@rockefeller.edu.

ORCID: UnJin Lee, 0000-0003-2424-650X.

Abstract:

A central open problem in the study of *de novo* gene origination is that the molecular mechanisms and functions driving the emergence of such evolutionarily young, *de novo* protein-coding genes remain poorly understood. Metal chelation, a simple, directly selectable activity that both requires no specific interaction partners and is also compatible with intrinsic disorder, is one possible function. This possibility was tested using sequence signatures in 7,849 transcriptionally supported, still-segregating *Drosophila melanogaster de novo* open reading frames. Interestingly, these new open reading frames (neORFs) are enriched for bis-histidine motifs at the metal-coordination-competent spacings H-x-H and H-x-x-x-H and show no enrichment at the incompatible even spacings relative to repeat-masked intergenic ORFs. Notably, these neORFs were also found to lack the C-x-x-C grammar of canonical metal-binding proteins. I report that this H-x-H bis-histidine signal is generated by translation of $(CA)_n$ microsatellites into His-Thr-His in *Drosophila melanogaster*, as evidenced by a CAC-codon bias within H-x-H motifs and a fourfold threonine enrichment at the central position. I propose that recurrent microsatellite expansion supplies *Drosophila* with a distributed, independently originated class of candidate metal-binding *de novo* peptides. I then trace one neORF (ZMEG) from conserved ancestral non-coding sequence to a transcribed, *melanogaster*-lineage open reading frame whose (CA)9-derived His run presents a candidate His32–His36 bis-histidine site. I propose that such *de novo* proteins may constitute a class of molecules united not by common descent but by their shared origin in evolvable repeat sequence, highlighting the importance of emergence bias in molecular evolution.

Significance Statement

Where do brand-new genes come from, and what do they do? Most genes are inherited and modified over eons, but some protein-coding genes arise de novo from previously non-coding DNA, and their functions have remained mysterious. Analyzing thousands of young Drosophila genes, I find that an abundant, highly mutable class of repetitive DNA, (CA)n microsatellites, repeatedly spells out histidine pairs capable of grabbing metal ions such as zinc. This links a simple, error-prone genomic feature to a concrete, selectable biochemical activity, and traces one such gene from non-coding ancestry to a candidate metal-binding peptide. The findings suggest evolution can invent biochemistry not by copying old genes but by exploiting the predictable emergence of repeat sequences.

Are de novo genes real?

While the majority of work on new gene origination, or the appearance of evolutionarily young genes, has been focused on duplicate gene evolution and gene family expansions (*1*), much effort has been made to both establish and validate the existence of genes that have arisen *de novo*, i.e. protein-coding genes that have arisen from ancestrally inactive genomic sequences (*2*). Early work focused particularly on understanding this phenomenon on both a species (*3*) and population (*4*) level within *Drosophila*. Despite lingering doubts, subsequent follow-up work has been performed in order to characterize the biophysical properties of such *de novo* genes (*5–7*), revealing common trends, such as high levels of intrinsic disorder and short ORF length (*7*) — effects that have generally been ascribed to pre-adaptation (*6*, *8*), and the ambush hypothesis (*9*). Despite this work, widespread acceptance of the existence of *de novo* genes has been hampered by the lack of ascribable function for these genes.

Prior efforts have attempted to explain the evolutionary forces driving the origination and retention of these genes, such as the proto-gene (*10*), transcription-first (*2*), and pre-adaptation models (*6*, *8*). However, such models are generally unsatisfying in that they do not directly tackle the question of function. One exception is the recently proposed cultivator model of *de novo* gene origination (*11*, *12*), which posits that the transcriptional activation of a new promoter is a crude but highly evolvable mechanism that selects for transcriptional activity not of the newly evolved *de novo* transcript, but of its neighboring genes through physical, distance-based effects such as transcriptional interference (*13–15*) or supercoiling mediated coupling (*16*). Indeed, such rapid evolvability has been shown to be critical in the understanding of the evolutionary trajectories of duplicate genes arising via some form of emergence bias (*17*), either through the innovation-amplification-divergence model (*18*, *19*), and/or the enhancer capture-divergence model (*20*). However, the cultivator model remains unsatisfying, as it can only provide an evolutionary mechanism describing the earliest stages of *de novo* gene origination, i.e. non-coding RNA (ncRNA), and remains unsatisfactory for explaining the persistence and function of protein-coding *de novo* genes.

The general acceptance of *de novo* origination has also been further hampered by key confounding technical factors found in many *de novo* gene studies, e.g. short peptide length (*7*), and the natural limits of homology identification in divergent species (*21*). Instead, putative *de novo* gene origination events are difficult to untangle from the duplication-based origination mechanisms underlying orphan genes (*22*) — duplicate genes that have diverged so rapidly from their parental gene that homology is impossible to detect (*21*). However, the recent explosion of high-quality genome annotations of a vast number of species has accelerated the study of *de novo* genes, providing increased support for claims of *de novo* origination beyond the absence of

detectable homology. For example, phylostratigraphic methodologies heavily rely on synteny to *positively* demonstrate the absence of transcriptional units in ancestral states (*5*, *10*, *23*). While such approaches do not necessarily exclude the possibility of duplication, either symmetrically in tandem or asymmetrically over larger distances that traverse across topologically associating domains (TADs), they do allow for the positive identification of an absence of a transcription unit *at an ancestral locus*. Other efforts have also utilized direct identification of ancestral sequence in combination with synteny to positively identify the lack of transcription via the usage of mock annotations in outgroup species (*12*). By using RNA-seq data from outgroup species, such an approach can definitively demonstrate that sequences ancestral to *de novo* genes (used here to include both protein-coding and ncRNA genes) were truly transcriptionally inactive (*12*). However, this approach has been used mostly to identify *de novo* non-coding RNAs, leaving a gap in our understanding of how *de novo* protein coding genes may arise naturally.

What do de novo proteins do?

A comprehensive analysis of protein function provides the following list of possible mechanisms driving *de novo* gene fixation. First is pre-adaptation, the claim that selection acts to minimize the genomic harm done by previously fixed *de novo* proteins subsequent to the fixation of a random protein-coding gene (*6*, *8*). Unlike the other putative functions listed here, this function — or lack thereof — is driven by negative selection against some unknown deleterious interaction. Another possibility is the appearance of a fortuitous but beneficial interaction with some number of pre-existing interaction partners that drives the new *de novo* protein to fixation. Possible sub-mechanisms include general protein-protein interaction via disordered regions (*24*), the refinement of regulatory function of pre-existing genes (*11*), or the titration of

thermodynamic properties of liquid-liquid phase separated condensates (LLPS). This last possibility also opens the door to more interesting possibilities, allowing for the potential of *de novo* proteins playing roles in cellular-scale phenomena either via direct interaction with condensates or surface adsorption (*25*). One concrete example would be similar in nature to the duplicate gene HP6/Umbrea, a known heterochromatin protein (*26*, *27*). Another potential mechanism would be membrane-based interactions (*28*). Whatever the case, protein-protein interactions often have strong sequence-specific signatures in the form of short linear motifs (SLiM)(*29*), such as in the case of the known PxVxL signature found in HP6/Umbrea (*26*).

One previously under-appreciated possibility is that *de novo* proteins may act as chelating agents, a function that is highly consistent with commonly observed properties of *de novo* genes such as intrinsic disorder. This is an attractive possibility, as chelation provides a simple, direct, selectable function that does not specifically require interaction between a nascent *de novo* protein and the rest of the proteome. While some degree of interaction is likely to occur between nascent *de novo* proteins and other proteins, variants producing such proteins are expected to be retained so long as the net selective effect of such incidental interactions remains positive. Further refinement or purging of such deleterious interactions could be explained by pre-adaptation (*8*), but the mechanism driving the original fixation events remains to be explained under that model.

Two primary signatures are known to be associated with chelation. First, pairs of cysteine residues in the configuration C-C or C-x-x-C are widely recurring motifs in metal- and disulfide-coordinating proteins (*30*). In *Drosophila*, proteins bearing these motifs, such as the cysteine-rich metallothioneins, are transcriptionally induced by the MTF-1 (metal-responsive transcription factor 1) pathway (*31*). Additionally, pairs of histidine molecules are also known motifs for

chelation due to histidine's partially protonated imidazole residue (*32*). The ordering for these histidines is critical for establishing metal-binding function, manifesting as H-x-H and H-x-x-x-H motifs, where x is an uncharged residue (*33*). Importantly, H-H and H-x-x-H motifs are disfavored for metal-binding function, as the imidazole residues chelate only when both residues converge on the same ion (*32*, *34*, *35*). Because histidine is itself the substrate of METTL9-family protein histidine methyltransferases, whose methylation can attenuate metal binding, any (CA)n-derived histidine site is in principle subject to post-translational control (*36*, *37*). In order to test for bis-histidine residue chelation as a mechanism driving *de novo* protein origination and retention, I utilized an existing dataset of high-confidence, population-level, new expressed ORFs (neORFs) to test for enrichment of histidine chelation-associated amino acid signatures in a panel of 7867 transcriptionally supported, young *de novo* proteins (neORFs) that are still segregating in *Drosophila melanogaster* (Materials and Methods)(*5*). Importantly, neORFs consist of the category of *de novo* ORFs that appear in fewer than 7 strains out of a total of 7 independent, iso-female lines of *D. melanogaster* (*5*). I then identify and analyze potential biochemical function of a candidate chelating neORF, which I term *zmeygorynych* (ZMEG), after the mythological three-headed dragon described in Eastern Slavic folklore.

Results

neORFs are enriched for H-x-H bis-histidine coordinate chemistry

To test for potential enrichment of bis-histidine-driven chelation in *de novo* proteins, the relative number of various H-(x)n-H configurations was calculated for three classes of validated, transcriptionally active genes: neORFs (N=7,849, $\ell_{0.5}$=64 aa), transcriptionally inactive, intergenic, repeat-masked ORFs (N=1,472,446, $\ell_{0.5}$=22 aa), and known metal-binding proteins

from UniProt (KW-0479)(N=856, $\ell_{0.5}$=538 aa). Specifically, the number of motifs of each category was calculated, then compared to the number of motifs expected by chance alone (Materials and Methods). In addition to a cyclic pattern of enrichment for odd n, consistent with bis-histidine chelation, the number of detected H-x-H patterns in neORFs was significantly enriched over the expectation from repeat-masked intergenic ORFs extracted from the *D. melanogaster* genome (Figure 1A). Additionally, no significant difference was found in the inert H-x-x-H configuration. The relative number of neORFs again increases in the H-x-x-x-H configuration, then continues to repeat in a cyclical manner, consistent with the geometry of bis-histidine chelation—though, as established below, this odd-spacing periodicity is a mechanical consequence of the underlying (CA)n reading frame rather than, in itself, evidence that chelation drives the origination or retention of neORFs in *Drosophila melanogaster*. Notably, the relative number of inter-histidine residues reached a maximum at n=3 (H-x-x-x-H) for metal binding proteins, consistent with the α-helix register. This reflects different motifs predominant in HExxH zincin proteases, including the gluzincins thermolysin and neprilysin, and the metzincins MMPs and ADAMs (*38*, *39*), and C2H2 zinc fingers, where both motifs coordinate the ion with two histidines one helical turn apart (*40*).

Interestingly, repeat-masked intergenic sequences showed elevated levels of H-x-H sequences in comparison to GC- and length-controlled random sequences, suggesting that intergenic sequences may be poised for generating H-x-H sequences when compared to random sequence (Figure 1B). One possibility is that such sequences arise from (CA)n dinucleotide repeats (see below). Additionally, another potential driving force may be recent pseudogenization of previously existing neORFs. The possibility of ancestral or residual protein-coding capacity is partially supported by the observation that the median length ($\ell_{0.5}$) of

intergenic ORFs increases from 22 aa to 40 aa if the intergenic ORF has ≥ 2 histidine units. This latent potential for H-x-H motif appearance is also elevated when compared to a length-matched proteome sample (N=7,849, $\ell_{0.5}$=64 aa), suggesting that the presence of H-x-H motifs is actively selected against and averaged over in proteins that do not bind to metals. The prevalence of H-x-H in neORFs is also notable, as there is no significant difference between neORFs and metal-binders specifically within H-x-H motifs (Figure 1B).

The preference for neORFs to show H-x-H motifs is striking in comparison to C-x-x-C chelation motifs. Here, random sequence, the length-matched proteome sample, intergenic sequence, and *de novo* neORFs appear to have no notable preference for C-x-x-C motifs. Alternatively, the C-x-x-C motif is a strong feature of known metal-binding proteins, showing a nearly 5x elevation of motif prevalence in comparison to all other categories (Figure 1C). This result suggests that the known class of metal-binding proteins have at least two distinct mechanisms by which they chelate metals, one that is shared with neORFs, and one exclusive to old, conserved metal-binding proteins (see "De novo genes as a potential zinc regulator"). An examination of both the raw incidence rate and the density of H-x-H vs C-x-x-C motifs further supports the conclusion that neORFs are preferentially enriched for bis-histidine coordinate chemistry, rather than C-x-x-C chelation. Specifically, while the incidence and density of these motifs in known metal-binders remains high, the length-matched proteome sample is depleted for H-x-H motifs in comparison to neORFs, this pattern is reversed for C-x-x-C motifs (Figure 1D-E). Interestingly, while the cluster incidence of H-x-H and C-x-x-C motifs is elevated for known metal-binding proteins in comparison to other gene classes, the density of these motifs is relatively low in metal-binding protein resulting from generally longer protein length (Figure 1E).

(CA)n dinucleotide repeats drive potential neORF bis-histidine coordinate chemistry

While the consequences of H-x-H appearance are consistent with a selectable function via chelation (see *Discussion*), the H-x-H motif is not among the top 15 differentially prevalent SLiM present in the Eukaryotic Linear Motif database (*29*)(Figure S1). Additionally, while neORFs are enriched for histidine residues when compared to the length-matched proteome sample, neORFs are even more significantly enriched for cysteine than for histidine, thus providing an inadequate explanation for the simultaneous prevalence of the H-x-H but not the C-x-x-C motif (Figure S2). One possibility is that H-x-H motifs arise via the activation and lengthening of ORFs containing (CA)n repeats, the most highly abundant dinucleotide repeat found in the *Drosophila melanogaster* genome (Table 1). Specifically, the sequence (CA)5 = CAC ACA CAC ANN, translates to His-Thr-His-(...). To test this hypothesis, I compared codon usage in the 2,036 histidines contained in H-x-H motifs to the 17,546 total histidines contained in neORFs, as histidine can be produced using either the CAC or CAT codon. This analysis revealed that histidines within H-x-H motifs were strongly biased towards utilizing the CAC codon over the CAT codon (56.9% vs 43.1%), consistent with (CA)n origination. This stands in contrast to codon usage across all neORF histidines, where codon usage was reversed (48.3% vs 51.7%)(two-sided permutation shuffle test, k=5000, $p = 2.0 \times 10^{-4}$, Figure 2A). To further understand this phenomenon, I examined the relative frequency of amino acid usage for the middle position of the H-x-H motif. Consistent with the (CA)n hypothesis, I found that threonine was the most prevalent amino acid utilized in the middle position. When compared to background, threonine produced a 4.2x elevation over expectation (two-sided nonparametric bootstrap)(k=5000, $p = 4.0 \times 10^{-4}$), suggesting that (CA)n dinucleotide repeats were the likeliest

source of H-x-H motifs in neORFs. Note that while RepeatMasker was applied prior to later analyses, most (CA)n repeats from n=7 to n=9 typically are not masked due to their short length.

A (CA)n-to-His-x-His spine converts microsatellite expansion into recurrent bis-histidine chemistry

The codon-usage and central-residue data point to a single, mechanistically explicit route from DNA to bis-histidine chemistry (Figure 3A). Read on the sense strand, (CA)n microsatellites present the repeating codons CAC·ACA·CAC·ACA…, translating as His-Thr-His-Thr…, where histidine residues fall at every second residue, with the intervening position fixed as threonine. Each additional CA unit that remains in frame thus appends one His-Thr pair and lengthens the run, whereas an out-of-frame unit shifts the reading frame and abolishes the motif, suggesting a single mechanism underlies the detected CAC-codon bias, the fourfold threonine enrichment at the central position, and the odd n-spacing (H-x-H, H-x-x-x-H) periodicity (Figures 1 and 2), without invoking selection at any individual locus. Note that translation of the complementary strand of the same tract reads TGT·GTG·TGT… = Cys-Val-Cys, a C-x-C register.

This pattern is readily apparent in the candidate chelation neORF *zmeygorynych* (ZMEG) (Figure 3B), which was identified as the second most highly expressing H-x(odd)-H neORF in a reanalysis of S2R+ expression data (see “Segregating neORFs are expressed in S2R+ cells”). Additionally, an analysis of ZMEG sequence shows that it originated by utilizing a His-Thr-His-Thr-His triple at positions His32/His34/His36 embedded within a five-histidine run (His27, His29, His32, His34, His36). Across the uniquely translated neORF set, 762 (9.7%) carry at least one H-x-H motif and 130 (1.7%) contain a poly-(H-x-H) His run of the kind produced by an in-

frame (CA)n tract (Figure 3C). ZMEG lies in this poly-(H-x-H) tail. As a result, *zmeygorynych,* a 55-codon open reading frame on chromosome arm 3R, was named after the mythological Eastern Slavic dragon Zmey Gorynych and chosen to reflect a number of traits unique to ZMEG: 1) ZMEG was originally identified in an iso-female line of Ukrainian origin (*5*), 2) dragons are known for their propensity to hoard metals, 3) like ZMEG's 3x H-x-H motif occurrences, Zmey Gorynych is typically described as possessing not one, but three heads, and 4) Zmey Gorynych's autochthonous origins is reminiscent of ZMEG's origination from ancestrally non-coding DNA.

Segregating neORFs are expressed in S2R+ cells

One necessary (but insufficient) prerequisite for presenting (CA)n microsatellites' latent bis-histidine chemistry to selection is the transcription of these open reading frames in relevant tissue types. Given the association between innate immune response and Zn/Mn chemistry (see "De novo genes as a potential zinc regulator"), I therefore quantified neORF expression in *Drosophila* S2R+ cells using a publicly available RNA-seq series in which cells were exposed to control, zinc, and manganese conditions (GSE99332; *41*). Two quality-control steps proved essential. First, a subset of neORFs annotated as intergenic and non-overlapping in fact overlapped annotated features or repeats, or lay in low-mappability regions where multi-mapping reads inflate apparent expression. Second, restriction to uniquely mapping reads and to loci passing a per-base mappability filter yielded 267 high-confidence, uniquely mappable, non-overlapping *de novo* ORFs (Materials and Methods)(Figure 4A, B).

This analysis revealed that *de novo* ORFs are transcribed, but at characteristically low levels (*5*). 31% (84/267) of these neORFs reached a mean expression of ≥0.1 TPM (transcripts per million), compared with 57% of annotated genes, while 69% of neORFs were effectively

undetected (<0.1 TPM) (Figure 4A). Only a handful of neORFs exceeded 1 TPM (Figure 4B). Expression was nonetheless constitutive rather than sporadic, as detected neORFs were expressed across replicates and conditions. Thus, the low absolute levels reflect weak, pervasive transcription rather than tight, condition-specific regulation. Notably, the most highly expressed neORFs were not H-x-H carriers (Table S1).

Interestingly, ZMEG is expressed at low levels (mean ≈0.9 TPM across conditions, ≈0.5 TPM in controls) and rises under manganese (≈1.5 TPM), an apparent ~2.8-fold uptick (DESeq2 maximum-likelihood $\log_2$ fold-change ≈ +1.5). Intriguingly, a second *de novo* neORF bearing an H-x-H motif was expressed at appreciably higher levels than ZMEG (≈4.8 TPM; Table S1), but lacked the signature H-T-H motif. Regardless, ZMEG is not significantly differentially expressed between conditions (Materials and Methods)(adjusted $p = 0.24$). More importantly, manganese is a broad, strong stressor in these data: 3,900 genes are differentially expressed under manganese versus 859 under zinc (roughly 4.5-fold more), with classic proteotoxic and metal-detoxification programs (Hsp70, metallothioneins) strongly induced. ZMEG's modest rise is therefore best read as part of a general stress-mobilization wave that broadens the pool of transcribed cryptic and *de novo* loci, not as a dedicated manganese response (see Discussion).

Consistent with cryptic rather than dedicated transcription, ZMEG's promoter-proximal region contains no canonical core promoter (no TATA box) and shows no enrichment of MTF-1 metal-response elements (MREs) above intergenic background, in contrast to the bona fide MTF-1 targets MtnA–E, whose promoters are strongly MRE-enriched (Materials and Methods). ZMEG is thus not an MTF-1/MRE-regulated locus, reinforcing the interpretation that its expression, along with its manganese-associated uptick, reflects pervasive, stress-amplified transcription rather than a locus-specific metal-response circuit.

In order to confirm that ZMEG's expressed sequence is genuine and matches its assembly, I reconstructed the ORF directly from the S2R+ reads, rather than from the reference alignment. This analysis recovered a sequence identical to the Grandchamp et al. (*5*) assembly, including the in-frame (CA)9 tract and the His32/34/36 triple. Additionally, because dinucleotide tracts are prone to alignment artifacts, tract length and reading frame were read by eye from the raw reads for ZMEG and its comparators throughout (Materials and Methods).

To determine whether ZMEG is representative or exceptional, I compared the neORF's two key properties ((CA)n-derived poly-(H-x-H) signature and stress-associated expression) across the *de novo* set, revealing a decoupling of these two properties. Indeed, the latent (CA)n-to-H-x-H reservoir is genomically pervasive. In the full uniquely translated neORF set (n = 7,849), 762 loci (9.7%) carry at least one H-x-H motif and 130 (1.7%) poly-(H-x-H) runs (Figure 3C). Expression, however, remains limiting, as further analysis of the 267 uniquely mappable, non-overlapping neORFs quantified in S2R+ cells shows that only 34 carry an H-x-H motif and 4 poly-(H-x-H) runs (Figure 4B). Of these four poly-(H-x-H) loci, ZMEG is the only one detectably expressed under metal stress (Table S1), suggesting that transcription serves as a significant barrier for the activation of individual neORFs.

ZMEG arose de novo in the melanogaster lineage from conserved ancestral sequence

To validate ZMEG's *de novo* origins and to reconstruct how its metal-relevant chemistry arose, the locus was traced across nine *Drosophila* species spanning roughly 50 My of divergence (Materials and Methods)(Figure 4C,D, Table S2). Critically, the ZMEG locus is syntenic in all nine genomes, where the region lies in the same intergenic interval flanked by the conserved single-copy anchors Gr93 and glec, and the upstream sequence is largely shared

317 across lineages. This observation confirms ZMEG's in-situ origination, positively excluding the
318 possibility that ZMEG is an orphan gene.

319 This analysis also revealed that the ATG and the immediately following N-terminal
320 codons are conserved across the *melanogaster* group and into *D. pseudoobscura* (~25 My),
321 which shares ZMEG's N-terminal peptide (MYNIFAHLTSG…). However, this ORF terminates
322 after 16 codons upstream of the (CA)n tract. Importantly, the syntenic locus itself is recoverable
323 in the most distant outgroup, *D. virilis* (~40–50 My), establishing a deep nucleotide-level anchor.
324 However, *D. virilis* translates only a short, divergent N-terminal peptide and encodes no
325 ancestral ZMEG protein. These results confirm that the full-length ORF and its bis-histidine
326 chemistry are derived solely within the *melanogaster* lineage in concordance with its
327 identification as a segregating neORF (*5*), and was not inherited from an ancient protein.

328 Critically, polarizing character states on the tree resolves a stepwise, microsatellite-driven
329 birth mechanism for ZMEG (Figure 4C). In the two most distant outgroups the ancestral ZMEG
330 reading frame terminates before the (CA)n tract. On the *melanogaster*-group stem, the ancestral
331 ZMEG reading frame extends past this ancestral early stop codon, opening and elongating the
332 nascent ORF. Interestingly, an in-frame (CA)n tract seeds a poly-histidine motif within the
333 *melanogaster* sub-group. Finally, the tract reaches (CA)9 in *D. melanogaster* and *D. mauritiana*,
334 translating to the full poly-(H-x-H) run with the His32/34/36 triple in a mature 55-codon ORF.
335 Crucially, this mature state is tract-length-dependent and hypervariable in *D. simulans* ((CA)6,
336 43 codons) and *D. sechellia* ((CA)7, 41 codons), as the same complex carries poly-histidine as a
337 shorter, frame-labile ORF. Alternatively, *D. yakuba* and *D. erecta* carry poly-histidine at
338 intermediate tract lengths.

The same lability is evident within *D. melanogaster*, where the tract segregates between 6 and 10 units across iso-female strains studied in (*5*), with in-frame deletions and frameshifting indels that lengthen, shorten, or disrupt the His run (Table S2). The (CA)n tract is thus a reservoir in flux rather than a fixed feature. Notably, ZMEG arises directly in-frame, with no clear evidence of an intervening functional-RNA stage, suggesting that its origin did not proceed through a cultivated ancestral noncoding-RNA intermediate (*11*).

ZMEG's His32–His36 pair is a candidate bis-histidine metal site

Three independent lines of evidence converge on a single candidate site formed by His32 and His36 (Figure 5). First, the above joint evolutionary-mechanistic analysis indicated that the His32/34/36 triple is the direct in-frame translation product of the (CA)9 tract, so the site sits exactly where the (CA)n spine predicts bis-histidine chemistry should appear. Second, geometry: an analysis of ZMEG's ESMFold-predicted structure (Materials and Methods)(*42*) models the His32 and His36 side chains residing 4.63 Å apart (imidazole-nitrogen separation), within the range spanned by two-histidine metal sites, and a modeled Zn^{2+} sits exactly at 2.31 Å from each coordinating nitrogen (Figure 5A). Third, a learned zinc-site predictor, Metal3D (Materials and Methods)(*43*), places its highest-probability voxel for ZMEG just 0.8 Å from the midpoint of the His32–His36 pair (Figure 5B). While such structure predictions do not fully establish the existence of metal bonding, the convergence of two structural predictors operating on the predicted fold, together with the independent evolutionary argument, yields a specific and falsifiable prediction: that ZMEG's putative chelating function lies in the His32–His36 pair's capacity for bis-histidine coordinate chemistry. Notably, unlike the idealized β-strand register of the (CA)n schematic (Figure 3A), ZMEG's predicted fold is helical, and the coordinating

362 histidines are related by an i,i+4 spacing (His32→His36) consistent with same-face presentation
363 on a helix (Figure 5A,C). The intervening residue at ZMEG's His-x-His positions is threonine,
364 within the substrate tolerance of the histidine methyltransferase METTL9. This raises the
365 possibility of a post-translational "off" switch (see *De novo genes as a potential zinc regulator*).
366
367 **Discussion**
368 This work connects, for the first time, a single *de novo* protein-coding gene with an
369 explicitly connected evolutionary and molecular mechanism. By bridging *de novo* gene
370 origination, zinc handling in *Drosophila*, and concrete biochemistry, this work introduces the
371 possibility of broader generalization of ligand-binding activity as a driver of *de novo* protein
372 origination. Importantly, the evolutionary (CA)n-to-H-x-H axis makes this connection tractable
373 because the same microsatellite that can be polarized on a phylogeny also specifies, through the
374 reading frame, the chemistry of the resulting peptide, so that evolutionary history and candidate
375 function are encoded in a single feature.
376
377 *De novo genes as a potential zinc regulator*
378 METTL9 is a methyltransferase that has been recently demonstrated to play a key part in
379 post-translational regulation of a broad range of substrate proteins via pervasive 1-
380 methylhistidine modification (*36*). METTL9 tolerates only a limited set of residues at the
381 intervening (x) position, including threonine, which is the most common residue for (x) found in
382 this analysis. As METTL9 does not specifically prefer threonine over other members of the A-N-
383 G-S-T motif (*36*), the enrichment of threonine in particular reflects (CA)n origins rather than co-
384 adaptation to METTL9. Notably, mammalian METTL9 methylates a histidine within the H-x-H

motif to install Nπ 1-methylhistidine, which acts as a post-translational off-switch by decreasing the metal-binding capability of target proteins. This has been shown to be the mechanism for calprotectin, the mammalian anti-bacterial S100A8/S100A9 heterodimer, of which the short, disordered S100A9 (MRP14) subunit is the methylation target (Figure 5C)(*44*). Structurally, calprotectin's manganese/zinc "site 2" is a hexahistidine cage assembled across both subunits (S100A8 His17/His27 and S100A9 His91/His95/His103/His105; , *45*), whose six histidine donors far exceed ZMEG's single candidate two-histidine pair. While calprotectin is conserved between humans and mice (*46*), no S100/calprotectin orthologue is annotated in *Drosophila*. Nπ-methylation of S100A9 His107 weakens its zinc affinity roughly sixfold (the dissociation constant rises from ~14.3 to ~90.9 μM) (*37*). Notably, upon S. aureus infection, the methylation of S100A9 His107 is reduced (*37*). This loss of methylation thus increases local sequestration of zinc, which starves S. aureus of metal ions and clears the infection. In line with this, METTL9-KO mice with constitutively unmethylated S100A9 were found to be more resistant to S. aureus infection, consistent with calprotectin's zinc-sequestration function as the causal mechanism of resistance (*37*).

In *Drosophila*, the METTL9 ortholog (mettl9/CG5339) is also an Nπ 1-methylhistidine His-methyltransferase, though this regiochemistry is inferred by orthology and has not been directly confirmed for the fly enzyme. Importantly, *Drosophila melanogaster* has a known metal response pathway, where copper, cadmium, silver, and zinc ingestion upregulates the MTF-1 transcriptional pathway (*47*, *48*) that regulates known copper transporters, e.g. Ctr1B and ATP7 (*49*), and known metallothioneins, e.g. MtnA–E (*50*). However, these metallothionein proteins are cysteine-rich, not histidine-rich. Additionally, unlike copper, cadmium, and silver, zinc only

407 weakly induces the metallothionein arm of the MTF-1 response, whereas the transporter arm,
408 e.g. ZnT35C, is robustly zinc-induced and MTF-1-dependent (*51*).

409 On the other hand, zinc regulation in *Drosophila melanogaster* remains relatively
410 understudied (*51*), where "little is still known about local redistribution and availability of Zn
411 within… the cell." (*52*). What is currently known about Zn response in fruit flies is mostly
412 limited to zinc transporters such as ZnT35C (*51*), not direct zinc chelators. Concurrently, while
413 zinc sequestration is well-characterized in the previously reported mouse infection model,
414 S100A9/calprotectin is a vertebrate-specific innovation as previously noted, with no identifiable
415 homolog in *Drosophila*. However, previous work has noted that fruit flies nonetheless mount a
416 Transferrin-1–mediated nutritional-immune response during infection (*53*). As of publication,
417 there are currently no known dedicated histidine-rich zinc-sequestration or buffering proteins in
418 *Drosophila*, suggesting that this key biological function may be served by the collective action of
419 an entire family of mutually unrelated, strain-specific neORF zinc chelators, rather than a more
420 traditionally ascribed duplication-based expansion of a known gene family.

421 Interestingly, recent work has demonstrated that *Drosophila* mettl9/CG5339 has altered
422 binding affinities for known human METTL9 targets. Combined with S100A9/calprotectin's
423 known *zinc*-binding function in mammals, and METTL9's known mechanism of H-x-H
424 methylation, my results suggest that *Drosophila melanogaster*'s zinc response may potentially be
425 acting through a separate, yet undercharacterized pathway, via the entire *class* of *de novo* H-x-H
426 proteins. Such interactions would be consistent with the documented *Drosophila*-specific
427 evolution of mettl9 binding, where substitutions in flanking and -x- residues were found to be
428 less tolerated than with human METTL9 (*54*). In that study, the authors demonstrated how
429 *Drosophila* mettl9 had increased stringency in activity, preferring to bind A, G, S, C, or T

residues, rather than N, Q, H, and P in the middle position. ZMEG provides a concrete candidate for this logic: its central (x) residue is threonine, which lies within the more stringent *D. melanogaster* METTL9 acceptor set (A/G/S/C/T; *54*), so Nπ-methylation of a histidine in its H-x-H motif would offer a plausible post-translational off-switch on a candidate *de novo* chelator, mirroring METTL9's demonstrated control of calprotectin's zinc affinity (*37*).

Promiscuous expression presents de novo H-x-H motifs to selection

For the latent metal-coordinating chemistry of a *de novo* H-x-H peptide to be visible to natural selection, its open reading frame must be transcribed and translated. Notably, however, neither step requires locus-specific regulation. Pervasive, low-level transcription of intergenic sequence is a general feature of eukaryotic genomes (*55*) and is the very property by which the neORFs analyzed here were identified as transcribed loci (*5*). Noncanonical and short ORFs are similarly engaged by ribosomes at appreciable rates (*56*, *57*). This promiscuous transcription and translation is expected to intensify under physiological stress as transcriptional and translational specificity relaxes. For example, previous work has shown that translational fidelity and start-site selection loosen during metal stress (*58*). Under such conditions, the pool of expressed *de novo* ORFs transiently broadens, sampling from a large reservoir of (CA)n-derived H-x-H peptides without any single locus having evolved a dedicated promoter. Consistent with this, *de novo* genes tend to originate in ancestrally closed chromatin and acquire expression through newly originated regulatory sequences rather than by co-opting preexisting open chromatin (*59*); pervasive, stress-amplified transcription thus offers a route by which such initially unregulated loci, including those bearing (CA)n-derived H-x-H motifs, can be presented to selection before dedicated de-repression and/or heterochromatin escape evolves.

This supplies the missing bridge between motif supply and selectable function. Specifically, (CA)n expansion can generate the latent capacity for H-x-H motif emergence in neORFs. Subsequently, promiscuous, stress-amplified expression then presents these nascent peptides and their bis-histidine coordinate chemistry to selection, which can retain and quantitatively tune those whose metal-coordinating activity proves beneficial. Critically, this route requires no sequence-specific regulatory innovation at any individual locus: a general, non-specific expression mechanism suffices to expose a compositional property (histidine spacing) to selection. This parallels the adaptive, evolutionarily tuned phase behavior of stress-responsive proteins, whose selectable function is a bulk sequence property realized without site-specific regulation (*60*, *61*), and the broader pre-adaptation view in which function can precede and drive the establishment of *de novo* genes (*6*).

Because it is a general property rather than a locus-specific adaptation, the transcriptional signature of this mechanism is expected to be diffuse and non-motif-specific. This is consistent with, though not resolved by, existing metal-challenge transcriptomic studies (see "Segregating neORFs are expressed in S2R+ cells"), and would be most directly testable by ribosome profiling under acute metal stress. However, it should be noted that pervasive, non-specific low levels of translation are abundant as well. As such, ribosome profiling still would not provide a "smoking gun" demonstration of the chelation hypothesis, as observed levels of non-specific translation would remain indistinguishable from noise. Consistent with this general, non-specific reading, the metal-challenge data analyzed here show a strongly asymmetric response, as manganese, a broad stressor, differentially regulates roughly 4.5-fold more genes than zinc (3,900 versus 859). Here, ZMEG's own manganese uptick is not individually significant

(adjusted p = 0.24), consistent with the diffuse, non-locus-specific signature this model predicts (Table S1).

De novo genes as an evolvable plasticity mechanism

Emergence bias is a critical feature underlying the dynamics of genome evolution (*17*). Specifically, natural selection can only select for features that are presented to it *in the order of emergence*. That is to say, natural selection cannot select for variants that have not yet appeared in a population. In species with high effective population sizes such as *Drosophila melanogaster*, this evolvability problem is amplified, as near-neutral variants are more effectively purged in such populations, preventing "tunneling" behavior between selectively beneficial states (*62*). This crucial detail is a defining aspect that separates the study of evolutionary genetics and the study of random peptides and/or directed evolution.

The emergence of H-x-H motifs, potentially in connection with the prevalence of (CA)n dinucleotides in *Drosophila*, offers the tantalizing possibility that a subset of *de novo* proteins have independently but collectively evolved to act in a module-like manner. Such collective transcription and translation could allow for the potential quantitative tuning of zinc response via alteration of the mutational target size of plastic response to metal consumption or bacterial infection. Such mutational target size, and thus the relative mutation rate of genic vs epigenetic responses to environmental shifts, has been shown to play a role in determining the long-term fate of plasticity (*63*). Notably, AC/(CA)n is the most abundant class of dinucleotide microsatellite in *Drosophila melanogaster*, accounting for ~53% of perfect dinucleotide microsatellite tracts (≥5 repeat units) on the euchromatic chromosome arms and ~54% by repeat length (Table 1), despite the genome's overall AT-richness (AT tracts form a substantial second

class at ~34%). Most (CA)n tracts are short (median 6 units), but a minority extend into a longer tail; tracts such as ZMEG's (CA)9 sit at this chelation-competent extreme in *Drosophila* (*64–67*), yielding approximately 2-3 His-Thr repeats. This well-documented prevalence of (CA)n microsatellite repeats in *Drosophila* thus produces His-Thr-His upon translation of the (+) strand, and Cys-Val-Cys (but not the chemically active Cys-Cys or Cys-x-x-Cys motifs) on the (-) strand . While the evolutionary forces underlying the limited appearance of longer (CA)n tracts remain a mystery (*68*), one outcome of these repeats is the likely rapid appearance of a wide class of intrinsically disordered proteins with the ability to sequester metal, related not by direct descent via duplication, but the repeat, independent origination of hundreds or thousands of the same protein *subunit* – again, arising *de novo* and not through duplication.

The possibility of such independent origination highlights the key role of evolvability in natural selection – this a mechanism is plausible only when such events are heavily biased to emerge (*17*). Similar effects arising from functional repeat units accelerating the evolution of genic solutions to fluctuating environments have been previously observed in bacterial promoters (*69*). As a dinucleotide repeat, it is important to note that expansion of this highly labile repeat is also likely to produce frameshifts within open reading frames, causing the H-x-H motif to be both highly evolvable but similarly easy to lose. As the mutation rate of such dinucleotide repeats is much higher than more commonly studied single nucleotide substitutions (*70–72*), this may allow H-x-H motifs from (CA)n repeats to potentially serve a secondary function as a favored launchpad for generating further protein coding diversity.

A latent reservoir and the limits of a genomic lens

Two important limitations must be acknowledged in framing this study. First, the metal-challenge series analyzed here may not reach the concentrations, duration, or cellular context of a physiologically relevant metal-stress episode, so the absence of a strong, ZMEG-specific transcriptional response in these data does not bound the reservoir's potential. Instead, it may simply indicate that the relevant regime was never sampled. Second, ZMEG is expressed at a level that would be diluted below detection in whole-animal RNA-seq, averaged across the many tissues in which it is silent. Notwithstanding ZMEG's original identification in (*5*), such expression became visible when expression was assayed in a defined cell population (S2R+ cells). The capacity for (CA)n-derived bis-histidine chemistry may sit in a latent, genomic reservoir (*73*), as demonstrated through a single ORF (ZMEG) that had previously gone unnoticed in published data sets. Here, I have shown that ZMEG's conversion into a transcribed, candidate metal-binding peptide is not merely feasible but, given the genomic prevalence of the underlying microsatellite, plausibly recurrent.

This same genomic reservoir is, paradoxically, extraordinarily difficult to interrogate through a genomic lens. If the mechanism proposed herein is a general, stress-amplified broadening of expression rather than a locus-specific circuit, then its signature in transcriptomic and/or translatomic data is expected to be indistinguishable from noise. Accordingly, sequencing cannot by itself distinguish a functionally engaged *de novo* chelator from incidental pervasive expression/translation. The two questions that would settle function are not genomic at all: does ZMEG actually bind metal inside the cell? and does genetically deactivating the (CA)n reservoir measurably shift intracellular metal levels? Neither is answerable by sequencing alone, highlighting the need for further experimentation. Beyond just a mere caveat, this work highlights, rather than weakens, the value of this work: the entire evolutionary-to-molecular

account assembled here, including origin, mechanism, and a testable candidate site, was recovered from a handful of publicly available datasets (a single focal species, eight outgroup genomes, and one previously published RNA-seq dataset), for a gene effectively invisible to standard genomic pipelines. Crucially, a motif-first search also would not have found ZMEG: one would have had to known the yet unannotated (ELM database)(Figure S1) H-x-H signature in advance, whereas here, the evolutionary analysis drove and surfaced the underlying chemistry. I highlight that this inversion is the broader thesis this work is meant to advance, demonstrating that comparative and population-level study of how sequences arise and change can serve not only as a post-hoc historical description of molecular history, but as a discovery tool for uncovering new biochemistry and biophysics.

Materials and Methods

New expressed open reading frame (neORF) dataset

7,894 neORFs from Grandchamp et al. (*5*) were retrieved from the associated data archive (Zenodo record 7322757). Because the published protein sequences appended stop-codon-derived nucleotides as C-terminal residues, neORFs were re-translated from their nucleotide entries in frame, terminating at the first in-frame stop. Re-translations were de-duplicated by amino-acid sequence, yielding 7,849 unique *D. melanogaster*-specific neORFs (median length $\ell_{0.5}$ = 64 aa). Following Grandchamp et al., neORFs are the *de novo* ORFs segregating in fewer than 7 of 7 independent iso-female lines, where neORFs are young, transcriptionally supported, population-level ORFs throughout.

Gene sets

neORFs were compared to four reference sets. (i) Intergenic, repeat-masked ORFs (N = 1,472,446; $\ell_{0.5}$ = 22 aa): six-frame, stop-to-stop ORFs (≥10 aa, no ambiguous residue, de-duplicated) extracted from the complement of (annotated genes ∪ RepeatMasker intervals) on the *D. melanogaster* autosome arms (dm6 + UCSC dm6 RepeatMasker track with manual coordinate verification). (ii) GC- and length-matched random ORFs: sequences generated at the intergenic GC content (0.413) and matched to the neORF length distribution, providing a genetic-code null that lacks genuine DNA structure. (iii) Length-matched proteome (N = 7,849): each neORF was matched with a nearest-length protein from the *D. melanogaster* proteome (FlyBase release dmel r6.61; dm6/BDGP6). (iv) Known metal-binding proteins (N = 856; $\ell_{0.5}$ = 538 aa): UniProt reviewed *D. melanogaster* proteins annotated with the Metal-binding keyword (KW-0479) (FlyBase release dmel r6.61, dm6/BDGP6)(*74*).

Bis-histidine spacing (order-fold)

For each set, H-$(x)_n$-H configurations (two histidines separated by n intervening residues, n = 1…6) were counted across all sequences. To isolate residue *order* from amino-acid *composition*, the observed pooled count was divided for each n by the mean count over a bootstrap of composition-preserving per-sequence scrambles (each sequence's own residues permuted; k = 3,000), giving a per-class "relative number" (order-fold). Significance between neORFs and intergenic ORFs was assessed with a two-sample, unpaired bootstrap test of the difference between pooled order-fold ratios. The identical procedure was applied to C-$(x)_n$-C configurations (C-x-x-C).

Motif incidence and density

For H-x-H and C-x-x-C motifs, incidence (fraction of proteins with ≥1 motif) and density (motifs per residue, i.e. length-normalized) were computed and reported across the random, length-matched-proteome sample, intergenic, neORF, and metal-binder sets. Because the sets differ ~25-fold in median length, the incidence/density split separates per-protein occurrence from per-length rate (Data S1).

Codon usage and the $(CA)_n$ mechanism

As histidine is encoded by CAC or CAT, codon usage was compared in the 2,036 histidines contained within H-x-H motifs to the 17,546 total histidines across neORFs, with significance assessed by a two-sided permutation shuffle test (k = 5,000). Relative amino-acid frequency at the central (x) position of H-x-H motifs was compared to background composition, with significance by a two-sided nonparametric bootstrap (k = 5,000). RepeatMasker was applied before ORF extraction, and short $(CA)_n$ tracts (n ≈ 7–9) were manually verified to escape masking due to length.

Short linear motifs and composition

neORFs and the length-matched proteome were scanned for the Eukaryotic Linear Motif (ELM) classes using the ELM regular-expression catalog (*29*), ranking classes by the difference in per-length density between neORFs and the proteome (Figure S1); H-x-H is not itself an ELM class. Per-class amino-acid composition (histidine, cysteine, aspartate, glutamate, and the full 20-residue profile) was computed across all sequence sets (Figure S2).

RNA-seq data and analysis

Expression was quantified from a publicly available *Drosophila* S2R+ RNA-seq series comprising control, zinc-, and manganese-exposed samples (GSE99332; *41*). Reads were aligned to the *D. melanogaster* genome (dm6/BDGP6)(*75*) with STAR (*76*), retaining only uniquely mapped reads, and were counted over neORF and annotated-gene intervals with featureCounts (*77*). Counts were then converted to TPM using ORF length. Two quality-control filters guarded against microsatellite- and repeat-driven mapping artifacts. First, neORFs nominally annotated as intergenic and non-overlapping were re-checked against gene and repeat annotations, and those overlapping annotated features or repeats were excluded. Second, per-base unique-mappability was computed across each locus, and low-mappability loci were removed. These filters yielded 267 uniquely mappable, non-overlapping *de novo* ORFs used for all expression estimates.

Differential expression between metal-exposed and control samples was assessed with DESeq2 (*78*) under a genotype-plus-condition design, with independent filtering and Benjamini–Hochberg correction; manganese-versus-control and zinc-versus-control were tested as separate contrasts, and the genome-wide breadth of differential expression was recorded for each. For ZMEG, the S2R+-expressed sequence was reconstructed directly from the mapped reads and compared to the reference assembly. Because dinucleotide microsatellites are prone to alignment-induced indel artifacts, the (CA)n tract length and the open reading frame were determined by manual inspection of the raw reads rather than from the reference alignment. Upstream regulatory content was assessed by scanning 1,200-bp windows around each candidate for MTF-1 metal-response elements (MRE core TGCRCNC, both strands) and canonical core-promoter motifs, calibrated against a null distribution of random intergenic windows and against the metallothionein (MtnA–E) promoters as positive controls.

Structure and metal-site prediction

The ZMEG open reading frame, and the S100A9 subunit of human calprotectin as a control, were folded with ESMFold (*42*) in the apo state (Figure S3); per-residue confidence is reported as pLDDT. Candidate metal sites and per-voxel zinc-binding probability were predicted with Metal3D (*43*), and the maximum predicted probability p(zinc) and the distance from the highest-probability voxel to the His32–His36 midpoint were recorded. Inter-histidine and metal–nitrogen distances were measured on the apo model with a Zn^{2+} placed for illustration only. Calprotectin was also analyzed identically.

Comparative synteny and de novo origin trajectory

Orthologous ZMEG loci were recovered in eight outgroup *Drosophila* genomes (*D. simulans*, *D. sechellia*, *D. mauritiana*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, and *D. virilis*) (NCBI assemblies GCF_016746395.2, GCF_004382195.2, GCF_004382145.1, GCF_016746365.2, GCA_047116545.1, GCF_017639315.1, GCF_009870125.1, and GCF_030788295.1, respectively) by anchored synteny to the flanking single-copy genes Gr93 and glec, followed by local alignment of the intervening region (BLAST, *79*, MAFFT, *80*). For each species, the ORF from the conserved ATG to the first in-frame stop, its length, histidine content, and (CA)n tract length were extracted directly from the assembly; because these genomes are single assemblies and microsatellite length is assembly-sensitive, non-reference tract lengths and their resulting frames are reported as provisional. Character states ((CA)n length, ORF length, and His-motif status) were polarized on the accepted *Drosophila* species topology to order the events shown in Figure 4C, and upstream sequence was aligned and polarized to distinguish ancestral from lineage-derived positions. Within *D. melanogaster*, the

(CA)n tract was additionally genotyped across the iso-female founder strains and its length and frame consequence (in-frame deletion versus frameshift) tabulated (Table S2).

Use of AI and generative AI software. The author used the generative AI large language model Claude Opus 4.8 (Anthropic) to assist with data retrieval, analysis, figure-generation code, and manuscript editing. The author reviewed, fact-checked, and takes full responsibility for all content.

Data Availability Statement

All analysis and figure-generation code, a single run-all wrapper, and intermediate result tables are provided in the accompanying code release (https://github.com/ulee-sciscripts/denovo_chelation); raw inputs are retrieved by the included fetch_data.sh from Zenodo record 7322757 (neORFs), FlyBase (proteome/annotation), UniProt (metal-binding proteins), and UCSC dm6 (genome, RepeatMasker), and the eight outgroup assemblies from NCBI (accessions in Materials and Methods). A permanently archived snapshot of the code release will be deposited at Zenodo [DOI to be assigned upon acceptance].

Acknowledgments

U.L. was supported by a National Science Foundation Postdoctoral Research Fellowship in Biology (NSF PRFB) award no. 2410289.

Competing Interests

The author declares no competing interest.

References

1. J. Zhang, Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18**, 292–298 (2003).

2. S. B. Van Oss, A.-R. Carvunis, De novo gene birth. *PLOS Genetics* **15**, e1008160 (2019).

3. M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, D. J. Begun, Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences* **103**, 9935–9939 (2006).

4. L. Zhao, P. Saelao, C. D. Jones, D. J. Begun, Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. *Science* **343**, 769–772 (2014).

5. A. Grandchamp, L. Kühl, M. Lebherz, K. Brüggemann, J. Parsch, E. Bornberg-Bauer, Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in Drosophila melanogaster. *Genome Research* **33**, 872–890 (2023).

6. B. A. Wilson, S. G. Foy, R. Neme, J. Masel, Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. *Nature Ecology & Evolution* **1**, 0146 (2017).

7. J. Peng, L. Zhao, The origin and structural evolution of de novo genes in Drosophila. *Nature Communications* **15**, 810 (2024).

8. B. A. Wilson, J. Masel, Putatively Noncoding Transcripts Show Extensive Association with Ribosomes. *Genome Biology and Evolution* **3**, 1245–1252 (2011).

9. H. Seligmann, D. D. Pollock, The Ambush Hypothesis: Hidden Stop Codons Prevent Off-Frame Gene Reading. *DNA and Cell Biology* **23**, 701–705 (2004).

10. A.-R. Carvunis, T. Rolland, I. Wapinski, M. A. Calderwood, M. A. Yildirim, N. Simonis, B. Charloteaux, C. A. Hidalgo, J. Barbette, B. Santhanam, G. A. Brar, J. S. Weissman, A. Regev, N. Thierry-Mieg, M. E. Cusick, M. Vidal, Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).

11. U. Lee, S. M. Mozeika, L. Zhao, A Synergistic, Cultivator Model of De Novo Gene Origination. *Genome Biology and Evolution* **16**, evae103 (2024).

12. U. Lee, C. Li, C. B. Langer, N. Svetec, L. Zhao, Comparative single-cell analysis of transcriptional bursting reveals the role of genome organization in de novo transcript origination. *Proceedings of the National Academy of Sciences* **122**, e2425618122 (2025).

13. T. Fukaya, B. Lim, M. Levine, Enhancer Control of Transcriptional Bursting. *Cell* **166**, 358–368 (2016).

14. S. H. Keller, H. Deng, B. Lim, Regulation of the dynamic RNA Pol II elongation rate in Drosophila embryos. *Cell Reports* **42**, 113225 (2023).

15. E. A. Leyes Porello, J. Wu, S. Fallacaro, E. Dreispiel Juan, R. T. Trudeau, J. A. Vidal, S. Lin, C. Nieto, A. Singh, M. Mir, B. Lim, Enhancer placement impacts transcriptional dynamics in Drosophila embryos. *Nature Communications* **17**, 6155 (2026).

16. C. P. Johnstone, K. E. Galloway, Supercoiling-mediated feedback rapidly couples and tunes transcription. *Cell Reports* **41**, 111492 (2022).

17. T. Fuqua, N. Vakirlis, Emergence biases in molecular evolution. *Genome Biology and Evolution*, evag195 (2026).

18. U. Bergthorsson, D. I. Andersson, J. R. Roth, Ohno's dilemma: Evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences* **104**, 17004–17009 (2007).

19. J. Näsvall, L. Sun, J. R. Roth, D. I. Andersson, Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence. *Science* **338**, 384–387 (2012).

20. U. Lee, D. Arsala, S. Xia, C. Li, M. Ali, N. Svetec, C. B. Langer, D. R. Sobreira, I. Eres, D. Sosa, J. Chen, L. Zhang, P. Reilly, A. Guzzetta, J. J. Emerson, P. Andolfatto, Q. Zhou, L. Zhao, M. Long, The three-dimensional genome drives the evolution of asymmetric gene duplicates via enhancer capture-divergence. *Science Advances* **10**, eadn6625 (2024).

21. C. M. Weisman, A. W. Murray, S. R. Eddy, Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLOS Biology* **18**, e3000862 (2020).

22. D. Tautz, T. Domazet-Lošo, The evolutionary origin of orphan genes. *Nature Reviews Genetics* **12**, 692–702 (2011).

23. N. Vakirlis, A.-R. Carvunis, A. McLysaght, Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* **9**, e53500 (2020).

24. J. Peng, L. Zhao, A predicted structural interactome reveals binding interference from intrinsically disordered regions. *PLOS Computational Biology* **22**, e1013899 (2026).

25. C. P. Brangwynne, C. R. Eckmann, D. S. Courson, A. Rybarska, C. Hoege, J. Gharakhani, F. Jülicher, A. A. Hyman, Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation. *Science* **324**, 1729–1732 (2009).

26. B. D. Ross, L. Rosin, A. W. Thomae, M. A. Hiatt, D. Vermaak, A. F. A. de la Cruz, A. Imhof, B. G. Mellone, H. S. Malik, Stepwise Evolution of Essential Centromere Function in a Drosophila Neogene. *Science* **340**, 1211–1214 (2013).

27. F. Greil, E. de Wit, H. J. Bussemaker, B. van Steensel, HP1 controls genomic targeting of four novel heterochromatin proteins in Drosophila. *The EMBO Journal* **26**, 741–751 (2007).

28. N. Vakirlis, O. Acar, B. Hsu, N. Castilho Coelho, S. B. Van Oss, A. Wacholder, K. Medetgul-Ernar, R. W. Bowman, C. P. Hines, J. Iannotta, S. B. Parikh, A. McLysaght, C. J. Camacho, A. F. O'Donnell, T. Ideker, A.-R. Carvunis, De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature Communications* **11**, 781 (2020).

29. M. Kumar, S. Michael, J. Alvarado-Valverde, A. Zeke, T. Lazar, J. Glavina, E. Nagy-Kanta, J. M. Donagh, Z. E. Kalman, S. Pascarelli, N. Palopoli, L. Dobson, C. F. Suarez, K. Van Roey, I. Krystkowiak, J. E. Griffin, A. Nagpal, R. Bhardwaj, F. Diella, B. Mészáros, K. Dean, N. E. Davey, R. Pancsa, L. B. Chemes, T. J. Gibson, ELM—the Eukaryotic Linear Motif resource—2024 update. *Nucleic Acids Research* **52**, D442–D455 (2024).

30. D. E. Fomenko, V. N. Gladyshev, Identity and Functions of CxxC-Derived Motifs. *Biochemistry* **42**, 11214–11225 (2003).

31. B. Zhang, D. Egli, O. Georgiev, W. Schaffner, The Drosophila Homolog of Mammalian Zinc Finger Factor MTF-1 Activates Transcription in Response to Heavy Metals. *Molecular and Cellular Biology* **21**, 4505–4514 (2001).

32. R. J. Sundberg, R. B. Martin, Interactions of histidine and other imidazole derivatives with transition metal ions in chemical and biological systems. *Chemical Reviews* **74**, 471–517 (1974).

33. B. L. Vallee, D. S. Auld, Zinc coordination, function, and structure of zinc enzymes and other proteins. *Biochemistry* **29**, 5647–5659 (1990).

34. I. L. Alberts, K. Nadassy, S. J. Wodak, Analysis of zinc binding sites in protein crystal structures. *Protein Science* **7**, 1700–1716 (1998).

35. S. Karlin, Z.-Y. Zhu, Classification of mononuclear zinc metal sites in protein structures. *Proceedings of the National Academy of Sciences* **94**, 14231–14236 (1997).

36. E. Davydova, T. Shimazu, M. K. Schuhmacher, M. E. Jakobsson, H. L. D. M. Willemen, T. Liu, A. Moen, A. Y. Y. Ho, J. Małecki, L. Schroer, R. Pinto, T. Suzuki, I. A. Grønsberg, Y. Sohtome, M. Akakabe, S. Weirich, M. Kikuchi, J. V. Olsen, N. Dohmae, T. Umehara, M. Sodeoka, V. Siino, M. A. McDonough, N. Eijkelkamp, C. J. Schofield, A. Jeltsch, Y. Shinkai, P. Ø. Falnes, The methyltransferase METTL9 mediates pervasive 1-methylhistidine modification in mammalian proteomes. *Nature Communications* **12**, 891 (2021).

37. D. Cao, M. Lv, C. Hu, S. Li, S. Wang, C. Xu, W. Pan, METTL9-catalyzed histidine methylation of S100A9 suppresses the anti-Staphylococcus aureus activity of neutrophils. *Protein & Cell* **15**, 223–229 (2024).

38. N. M. Hooper, Families of zinc metalloproteases. *FEBS Letters* **354**, 1–6 (1994).

39. W. Bode, F.-X. Gomis-Rüth, W. Stöckler, Astacins, serralysins, snake venom and matrix metalloproteinases exhibit identical zinc-binding environments (HEXXHXXGXXH and

Met-turn) and topologies and should be grouped into a common family, the "metzincins." *FEBS Letters* **331**, 134–140 (1993).

40. N. P. Pavletich, C. O. Pabo, Zinc Finger-DNA Recognition: Crystal Structure of a Zif268-DNA Complex at 2.1 Å. *Science* **252**, 809–817 (1991).

41. S. E. Mohr, K. Rudd, Y. Hu, W. R. Song, Q. Gilly, M. Buckner, B. E. Housden, C. Kelley, J. Zirin, R. Tao, G. Amador, K. Sierzputowska, A. Comjean, N. Perrimon, Zinc Detoxification: A Functional Genomics and Transcriptomics Analysis in Drosophila melanogaster Cultured Cells. *G3 Genes|Genomes|Genetics* **8**, 631–641 (2018).

42. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

43. S. L. Dürr, A. Levy, U. Rothlisberger, Metal3D: a general deep learning framework for accurate metal ion location prediction in proteins. *Nature Communications* **14**, 2713 (2023).

44. H. Daitoku, M. Someya, K. Kako, T. Hayashi, T. Tajima, H. Haruki, N. Sekiguchi, T. Uetake, Y. Akimoto, A. Fukamizu, siRNA screening identifies METTL9 as a histidine Nπ-methyltransferase that targets the proinflammatory protein S100A9. *Journal of Biological Chemistry* **297**, 101230 (2021).

45. S. M. Damo, T. E. Kehl-Fie, N. Sugitani, M. E. Holt, S. Rathi, W. J. Murphy, Y. Zhang, C. Betz, L. Hench, G. Fritz, E. P. Skaar, W. J. Chazin, Molecular basis for manganese sequestration by calprotectin and roles in the innate immune response to invading bacterial pathogens. *Proceedings of the National Academy of Sciences* **110**, 3841–3846 (2013).

46. I. Marenholz, C. W. Heizmann, G. Fritz, S100 proteins in mouse and man: from evolution to function and pathology (including an update of the nomenclature). *Biochemical and Biophysical Research Communications* **322**, 1111–1122 (2004).

47. V. Günther, U. Lindert, W. Schaffner, The taste of heavy metals: Gene regulation by MTF-1. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1823**, 1416–1425 (2012).

48. D. Egli, A. Selvaraj, H. Yepiskoposyan, B. Zhang, E. Hafen, O. Georgiev, W. Schaffner, Knockout of "metal-responsive transcription factor" MTF-1 in Drosophila by homologous recombination reveals its central role in heavy metal homeostasis. *The EMBO Journal* **22**, 100–108 (2003).

49. A. Selvaraj, K. Balamurugan, H. Yepiskoposyan, H. Zhou, D. Egli, O. Georgiev, D. J. Thiele, W. Schaffner, Metal-responsive transcription factor (MTF-1) handles both extremes, copper load and copper starvation, by activating different genes. *Genes & Development* **19**, 891–896 (2005).

50. D. Egli, H. Yepiskoposyan, A. Selvaraj, K. Balamurugan, R. Rajaram, A. Simons, G. Multhaup, S. Mettler, A. Vardanyan, O. Georgiev, W. Schaffner, A Family Knockout of All Four Drosophila Metallothioneins Reveals a Central Role in Copper Homeostasis and Detoxification. *Molecular and Cellular Biology* **26**, 2286–2296 (2006).

51. H. Yepiskoposyan, D. Egli, T. Fergestad, A. Selvaraj, C. Treiber, G. Multhaup, O. Georgiev, W. Schaffner, Transcriptome response to heavy metal stress in Drosophila reveals a new zinc transporter that confers resistance to zinc. *Nucleic Acids Research* **34**, 4866–4877 (2006).

52. J. A. Navarro, S. Schneuwly, Copper and Zinc Homeostasis: Lessons from Drosophila melanogaster. *Frontiers in Genetics* **8**, 223 (2017).

53. I. Iatsenko, A. Marra, J.-P. Boquete, J. Peña, B. Lemaitre, Iron sequestration by transferrin 1 mediates nutritional immunity in Drosophila melanogaster. *Proceedings of the National Academy of Sciences* **117**, 7317–7325 (2020).

54. L. Schroer, S. Weirich, M. Hammerstad, H.-P. Hersleth, I. A. Grønsberg, L. Hagen, G. Slupphaug, J. M. Malecki, A. Jeltsch, P. Ø. Falnes, E. Davydova, Orthologues of the human protein histidine methyltransferase METTL9 display distinct substrate specificities. *Journal of Biological Chemistry* **301**, 110318 (2025).

55. K. Struhl, Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural & Molecular Biology* **14**, 103–105 (2007).

56. J. Ruiz-Orera, X. Messeguer, J. A. Subirana, M. M. Albà, Long non-coding RNAs as a source of new peptides. *eLife* **3**, e03523 (2014).

57. N. T. Ingolia, G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. S. Talhouarne, S. E. Jackson, M. R. Wills, J. S. Weissman, Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Reports* **8**, 1365–1379 (2014).

58. K. Mohler, M. Ibba, Translational fidelity and mistranslation in the cellular response to stress. *Nature Microbiology* **2**, 17117 (2017).

59. L. K. Blair, J. M. Cridland, D. J. Begun, A. Kopp, De novo genes originate in regions of ancestrally closed chromatin. *Genetics*, iyag178 (2026).

60. J. A. Riback, C. D. Katanski, J. L. Kear-Scott, E. V. Pilipenko, A. E. Rojek, T. R. Sosnick, D. A. Drummond, Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell* **168**, 1028-1040.e19 (2017).

61. E. W. J. Wallace, J. L. Kear-Scott, E. V. Pilipenko, M. H. Schwartz, P. R. Laskowski, A. E. Rojek, C. D. Katanski, J. A. Riback, M. F. Dion, A. M. Franks, E. M. Airoldi, T. Pan, B. A. Budnik, D. A. Drummond, Reversible, Specific, Active Aggregates of Endogenous Proteins Assemble upon Heat Stress. *Cell* **162**, 1286–1298 (2015).

62. D. B. Weissman, M. M. Desai, D. S. Fisher, M. W. Feldman, The rate at which asexual populations cross fitness valleys. *Theoretical Population Biology* **75**, 286–300 (2009).

63. U. Lee, E. N. Mortola, E. Kim, M. Long, Evolution and maintenance of phenotypic plasticity. *Biosystems* **222**, 104791 (2022).

64. D. Bachtrog, S. Weiss, B. Zangerl, G. Brem, C. Schlötterer, Distribution of dinucleotide microsatellites in the Drosophila melanogaster genome. *Molecular Biology and Evolution* **16**, 602–610 (1999).

65. C. L. Ross, K. A. Dyer, T. Erez, S. J. Miller, J. Jaenike, T. A. Markow, Rapid Divergence of Microsatellite Abundance Among Species of Drosophila. *Molecular Biology and Evolution* **20**, 1143–1157 (2003).

66. M. V. Katti, P. K. Ranjekar, V. S. Gupta, Differential Distribution of Simple Sequence Repeats in Eukaryotic Genome Sequences. *Molecular Biology and Evolution* **18**, 1161–1167 (2001).

67. G. Tóth, Z. Gáspári, J. Jurka, Microsatellites in Different Eukaryotic Genomes: Survey and Analysis. *Genome Research* **10**, 967–981 (2000).

68. S. Kruglyak, R. T. Durrett, M. D. Schug, C. F. Aquadro, Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences* **95**, 10774–10778 (1998).

69. M. Barnett, L. Meister, P. B. Rainey, Experimental evolution of evolvability. *Science* **387**, eadr2756 (2025).

70. M. D. Schug, T. F. C. Mackay, C. F. Aquadro, Low mutation rates of microsatellite loci in Drosophila melanogaster. *Nature Genetics* **15**, 99–102 (1997).

71. M. D. Schug, C. M. Hutter, K. A. Wetterstrand, M. S. Gaudette, T. F. Mackay, C. F. Aquadro, The mutation rates of di-, tri- and tetranucleotide repeats in Drosophila melanogaster. *Molecular Biology and Evolution* **15**, 1751–1760 (1998).

72. B. Harr, C. Schlötterer, Long Microsatellite Alleles inDrosophila melanogasterHave a Downward Mutation Bias and Short Persistence Times, Which Cause Their Genome-Wide Underrepresentation. *Genetics* **155**, 1213–1220 (2000).

73. S. L. Rutherford, S. Lindquist, Hsp90 as a capacitor for morphological evolution. *Nature* **396**, 336–342 (1998).

74. L. S. Gramates, J. Agapite, H. Attrill, B. R. Calvi, M. A. Crosby, G. dos Santos, J. L. Goodman, D. Goutte-Gattat, V. K. Jenkins, T. Kaufman, A. Larkin, B. B. Matthews, G. Millburn, V. B. Strelets, the FlyBase Consortium, N. Perrimon, S. R. Gelbart, J. Agapite, K. Broll, L. Crosby, G. dos Santos, K. Falls, L. S. Gramates, V. Jenkins, I. Longden, B. Matthews, J. Seme, C. J. Tabone, P. Zhou, M. Zytkovicz, N. Brown, G. Antonazzo, H. Attrill, P. Garapati, D. Goutte-Gattat, A. Larkin, S. Marygold, A. McLachlan, G. Millburn,

A. Öztürk-Çolak, C. Pilgrim, V. Trovisco, B. Calvi, T. Kaufman, J. Goodman, P. Krishna, V. Strelets, J. Thurmond, R. Cripps, T. Lovato, FlyBase: a guided tour of highlighted features. *Genetics* **220**, iyac035 (2022).

75. R. A. Hoskins, J. W. Carlson, K. H. Wan, S. Park, I. Mendez, S. E. Galle, B. W. Booth, B. D. Pfeiffer, R. A. George, R. Svirskas, M. Krzywinski, J. Schein, M. C. Accardo, E. Damia, G. Messina, M. Méndez-Lago, B. de Pablos, O. V. Demakova, E. N. Andreyeva, L. V. Boldyreva, M. Marra, A. B. Carvalho, P. Dimitri, A. Villasante, I. F. Zhimulev, G. M. Rubin, G. H. Karpen, S. E. Celniker, The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Research* **25**, 445–458 (2015).

76. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

77. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

78. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

79. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

80. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).

Figures & Tables

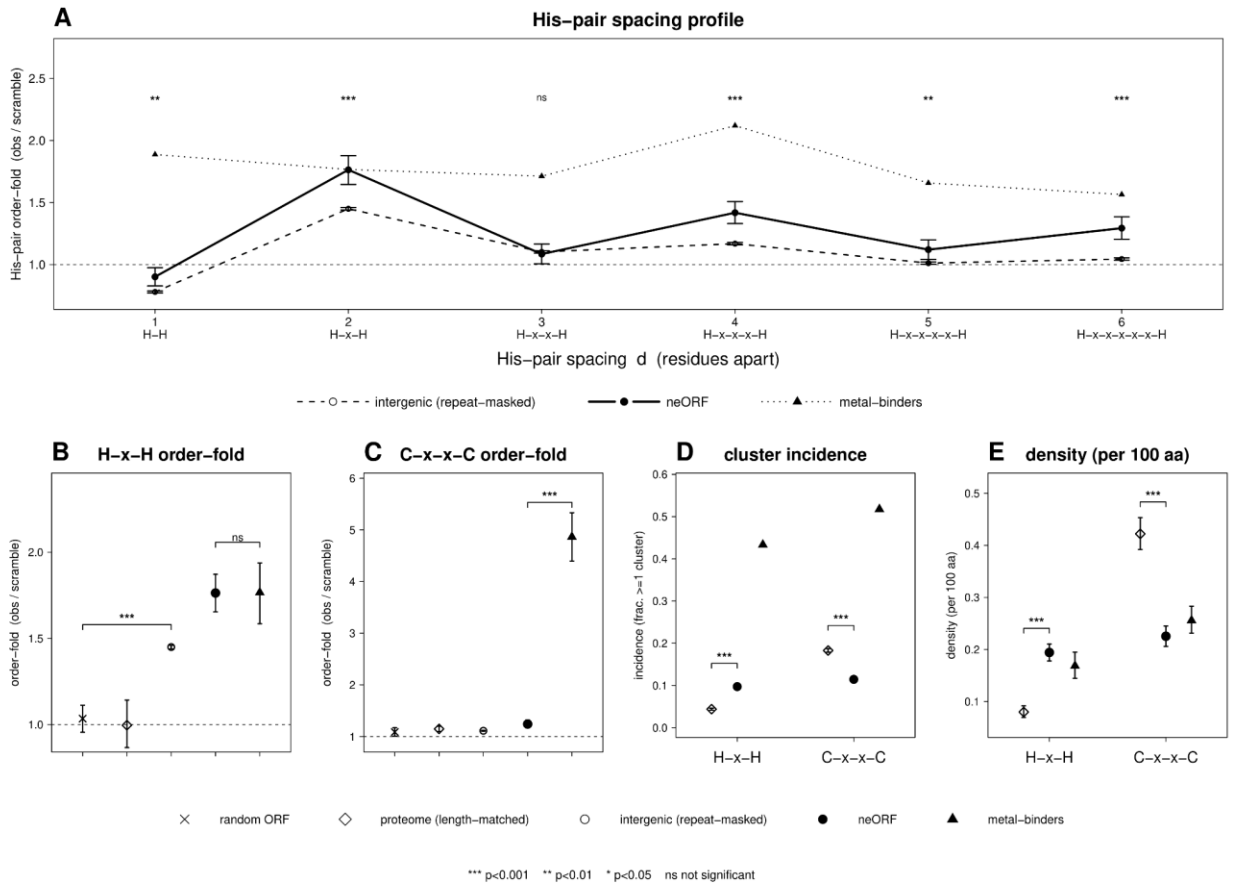


Figure 1: neORFs are enriched for H-x-H configurations but depleted of C-x-x-C configurations. (A) Relative number of H-(x)n-H configurations by gene class: neORFs (N=7,849), randomly selected intergenic, repeat-masked ORFs (N=1,472,446), and known metal-binding proteins (N=856). Raw numbers of configuration counts were divided by a bootstrapped distribution (k=3000) for each gene class. P-values were calculated between neORFs and intergenic ORFs using a two-sample, unpaired bootstrap test of difference between pooled order-fold ratios. Relative number of (B) H-x-H and (C) C-x-x-C configurations, and (D) cluster incidence and (E) density for H-x-H and C-x-x-C motifs in GC- and length-matched random ORFs, the length-matched *Drosophila* proteome, intergenic repeat-masked ORFs, neORFs, and known metal-binding proteins is shown.

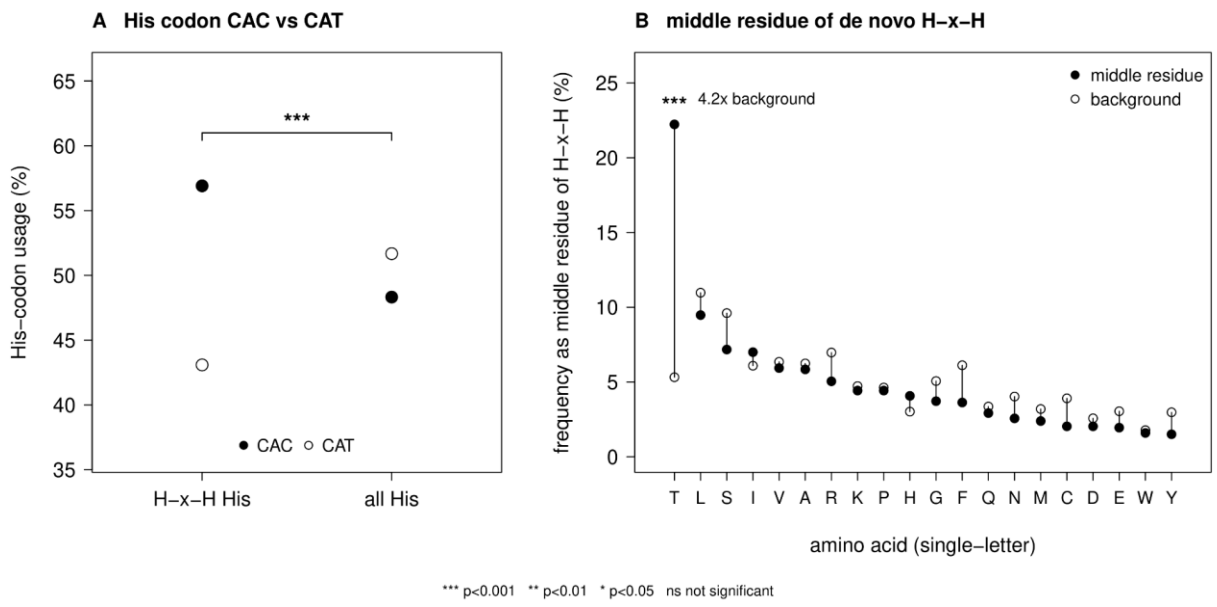
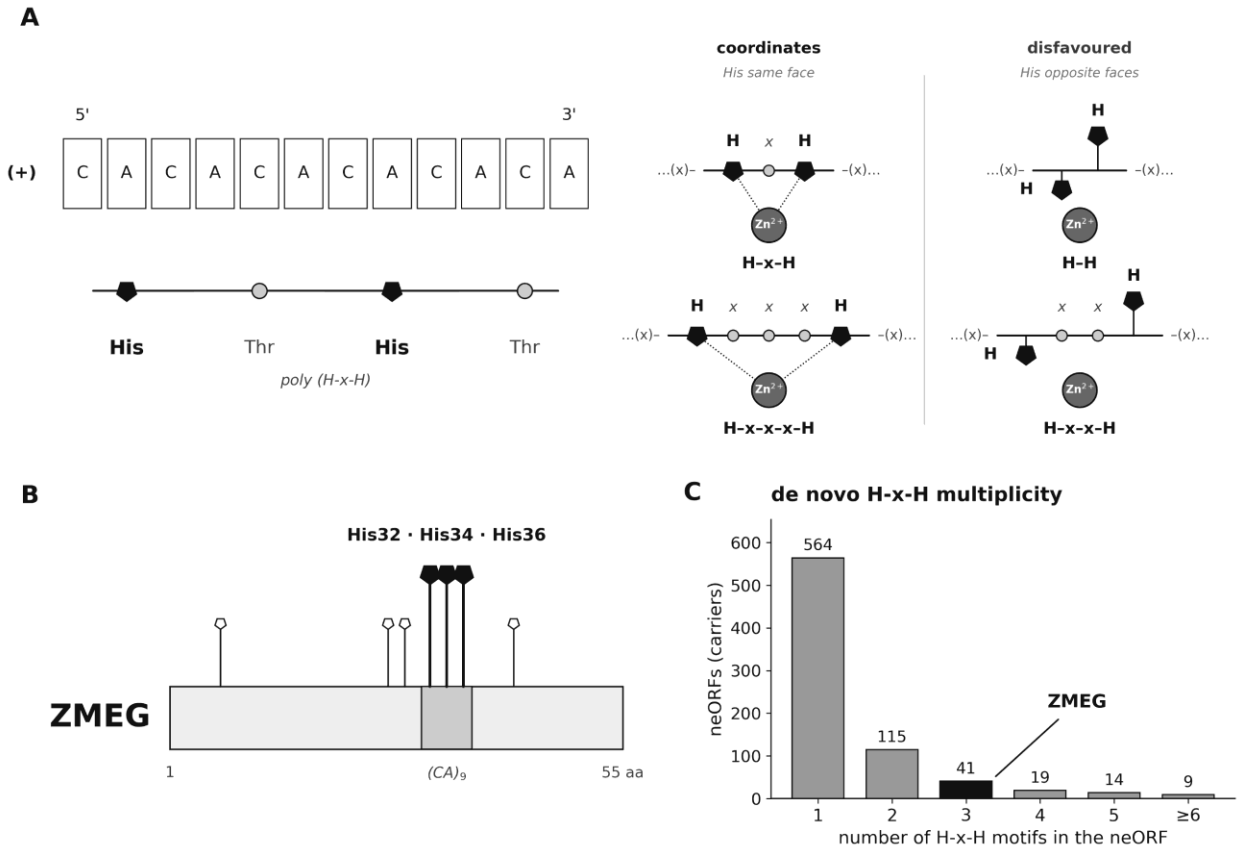


Figure 2: (CA)n dinucleotide repeats alter codon usage and amino acid positioning. (A) Histidine codon usage in both H-x-H motifs and all histidine residues found in *Drosophila* neORFs. Notably, histidines in H-x-H configurations are highly biased towards CAC usage. This trend is reversed when considering the set of all histidines found in neORFs. (B) Amino acid usage for middle residues (x) in bis-histidine motifs (H-x-H). The 4.2x elevation of Thr usage in the middle position is highly suggestive of (CA)n origins.



940

941 **Figure 3: (CA)n microsatellite origins for ZMEG.** (A) Illustration of (CA)n translation and

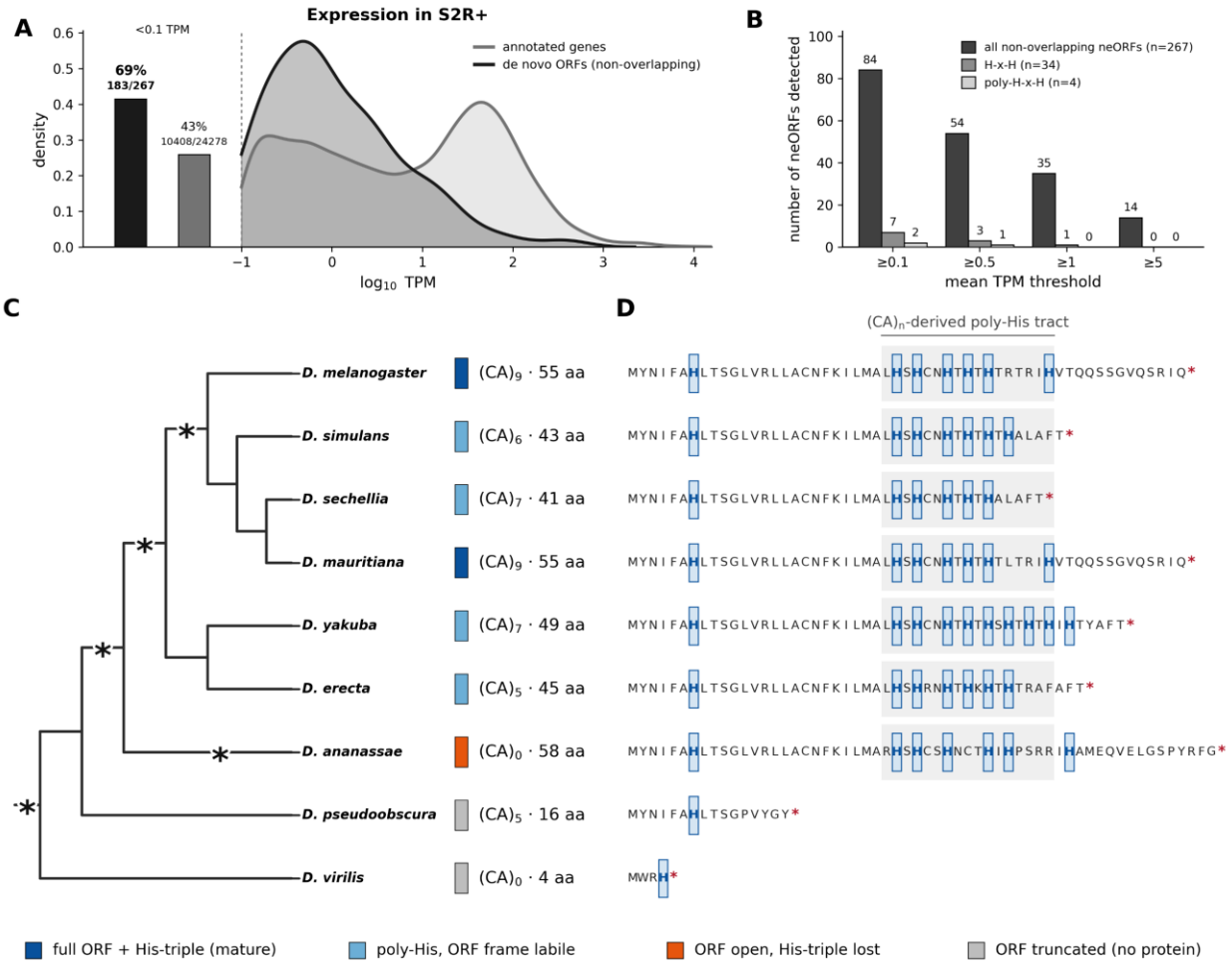
942 cartoon of bis-histidine coordinate chemistry. (B) A genomic (CA)9 tract translates in frame to

943 the His32/His34/His36 triple within a five-histidine run (His27–His36) contained in ZMEG. (C)

944 H-x-H motif multiplicity across the uniquely translated neORF set (n = 7,849), with ZMEG

945 indicated, demonstrates *de novo* proteins' latent capacity for bis-histidine coordinate chemistry.

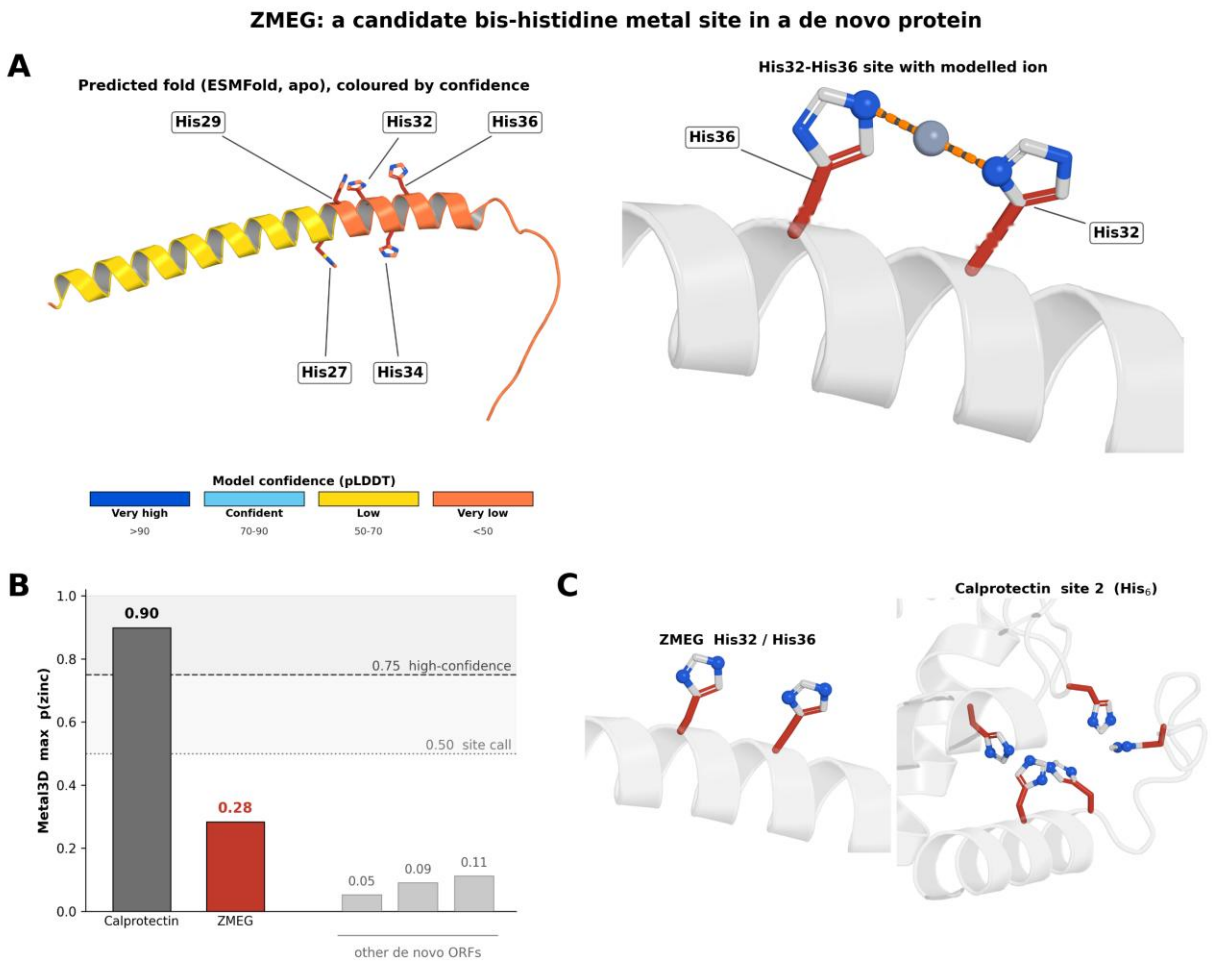
946



947

948 **Figure 4: *De novo* neORF transcription and the origins of ZMEG.** (A) Expression (mean
949 TPM across all conditions) of uniquely mappable, non-overlapping *de novo* ORFs (n = 267)
950 versus annotated genes in S2R+ cells. Left bars provide the fraction of undetected transcripts
951 (<0.1 TPM). (B) Number of *de novo* ORFs transcribed at increasing mean-TPM thresholds for
952 all neORFs and for the H-x-H (n = 34) and poly-(H-x-H) (n = 4) subsets. (C) *De novo* origin and
953 activation trajectory of ZMEG across nine *Drosophila* species (branch lengths not to scale),
954 including (CA)n tract length, ORF length to the first stop, and His-motif status mark. Asterisks
955 indicate polarized evolutionary events described in the main text. (D) ORF translations (ATG to

956 first stop) aligned across all nine species: histidines are boxed and (CA)n-derived poly-histidine
957 tracts are shaded.
958



959

960 **Figure 5: A candidate bis-histidine metal site in ZMEG.** (A) ZMEG predicted fold
961 (ESMFold, apo; colored by pLDDT model confidence) with the His27/29/32/34/36 run
962 indicated, and a close-up of the His32–His36 site with an illustrative modeled Zn^{2+} (apo
963 prediction; imidazole-nitrogen separation 4.63 Å, metal–nitrogen 2.31 Å). (B) Maximum
964 predicted zinc-site probability (Metal3D): ZMEG (0.28) versus a calprotectin S100A8/S100A9
965 positive control (0.90) and other *de novo* ORFs (0.05–0.11); dashed lines mark the tool's 0.5 and
966 0.75 confidence thresholds. For ZMEG the highest-probability voxel lies 0.8 Å from the His32–
967 His36 midpoint. (C) Structural comparison of ZMEG's loose two-histidine pair with
968 calprotectin's six-histidine metal cage.

Dinucleotide class	Tracts (n)	% of tracts	Repeat length (bp)	% by repeat bp
AC / (CA)n	11,323	52.9%	159,936	54.0%
AG	2,726	12.7%	38,610	13.0%
AT	7,271	34.0%	96,830	32.7%
CG	70	0.3%	728	0.2%
All classes	*21,390*	*100%*	*296,104*	*100%*

Table 1: Abundance of perfect dinucleotide microsatellite tracts by class in the *D. melanogaster* euchromatic genome. Perfect dinucleotide tracts (≥5 repeat units, ≥10 bp) were enumerated on chromosome arms 2L, 2R, 3L, and 3R of the UCSC dm6 assembly by our own reproducible census (see Methods; dinuc_census). Abundance is reported both by tract number and by total repeat length (bp). AC/(CA)n is the most abundant class by both metrics (~1.6× the next class, AT), consistent with the AT-rich composition of the genome. Literature surveys of *Drosophila* microsatellites report the same qualitative dominance of AC/(CA)n.

Supplementary Material

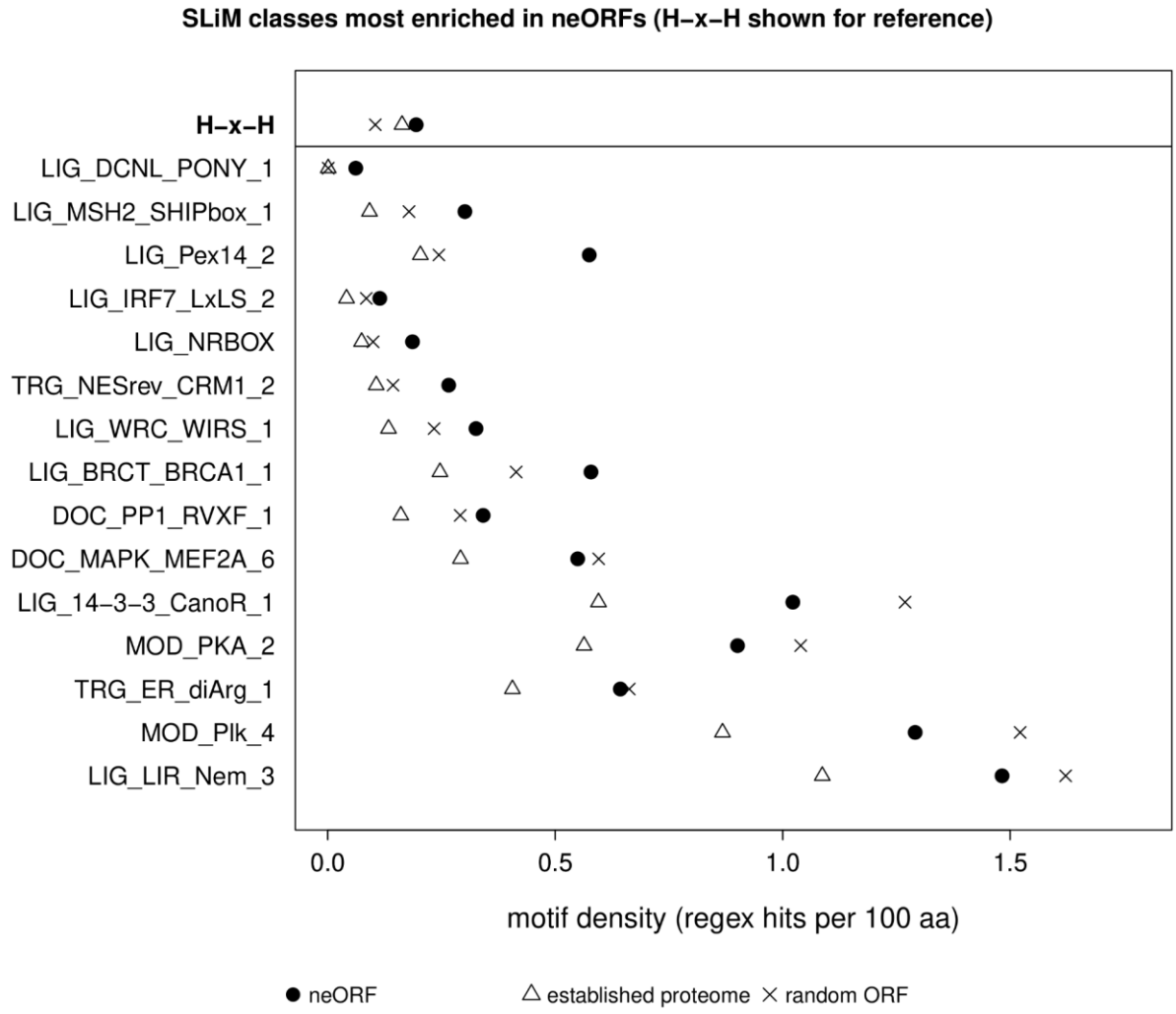
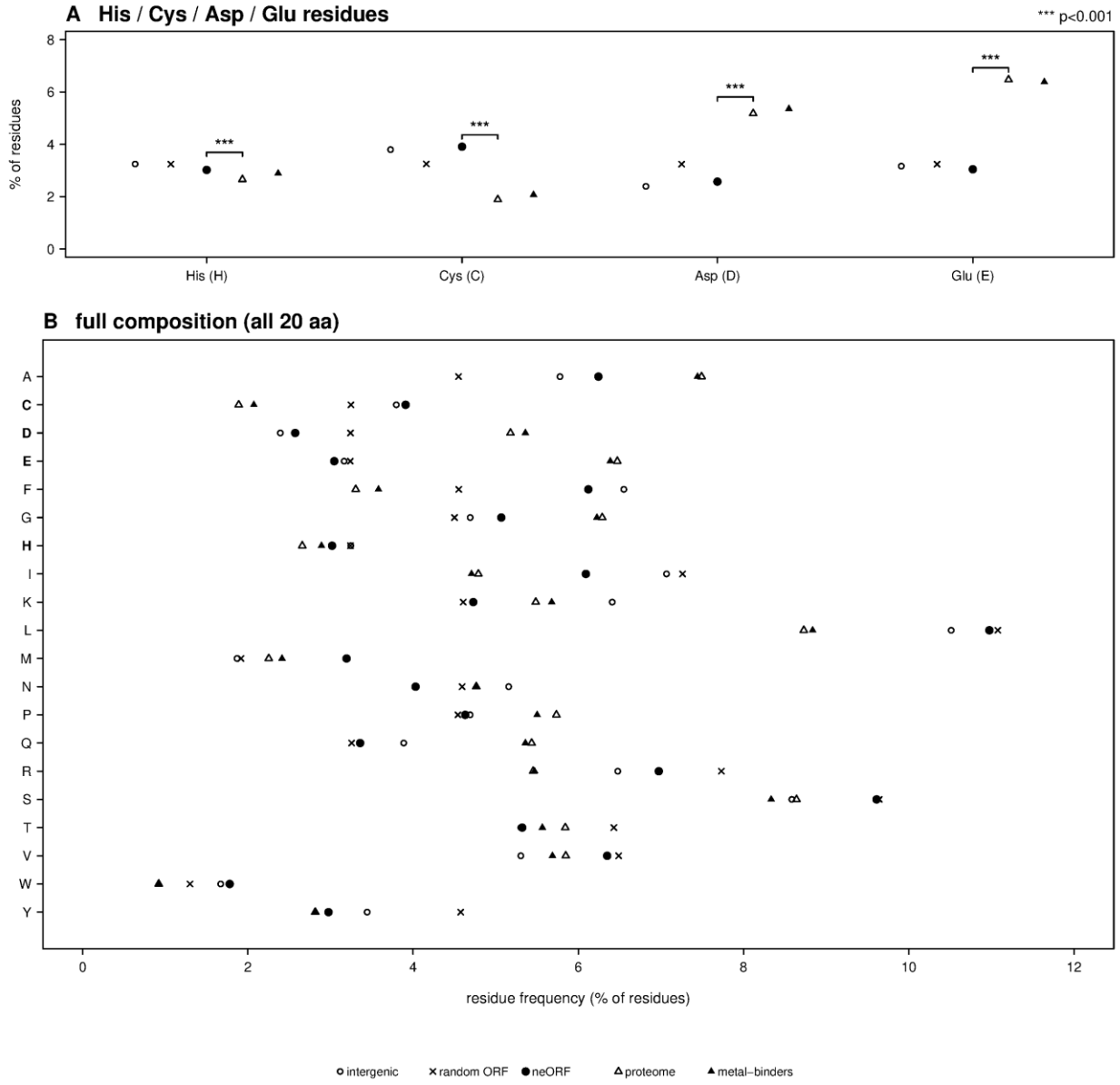


Figure S1. Top 15 most enriched SLiM in neORFs and H-x-H motif prevalence. Motif density for SLiM (ELM catalog) showing largest difference between neORFs and a length-matched proteome sample. Top row shows H-x-H, which was not present in ELM SLiM.



983

984 **Figure S2.** Histidine is not particularly enriched or prevalent in neORFs. Percentage of residues

985 calculated for (A) histidine, cysteine, aspartic acid, glutamic acid, and (B) all amino acids across

986 all gene classes.

987

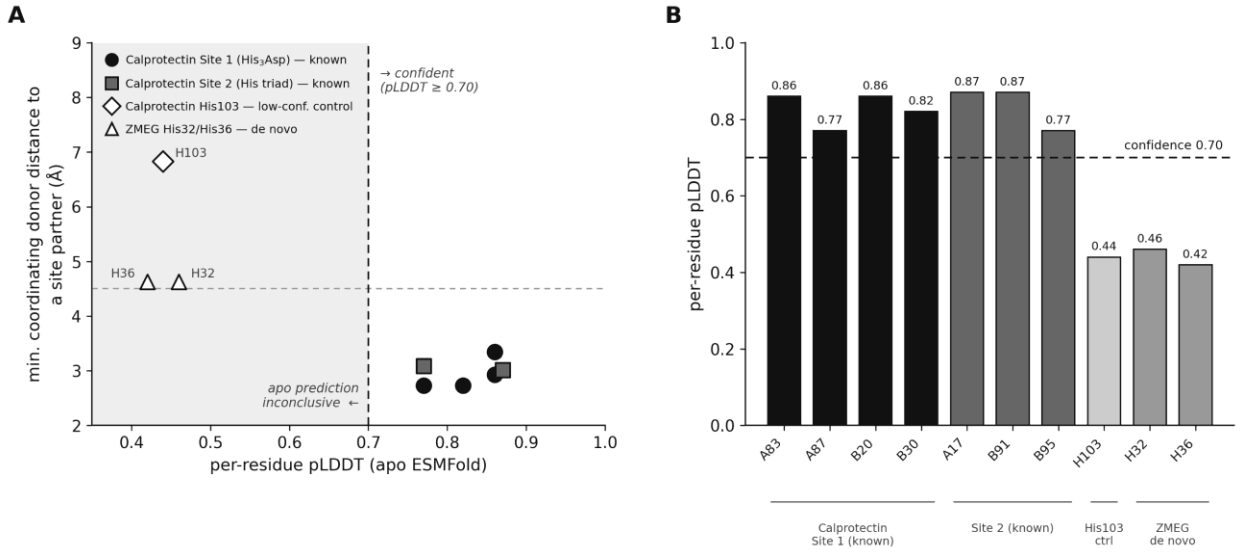


Figure S3. Apo structure prediction does not, on its own, recover calprotectin's metal site or known coordinating geometry for the S100A9 subunit of calprotectin, whose zinc/manganese site is experimentally established.

'TableS1_neORF_expression_267.csv'

Table S1. Per-replicate expression (TPM) of all 267 uniquely mappable, non-overlapping de novo neORFs in S2R+ cells across control, zinc, and manganese conditions and three genotypes, with H-x-H motif content and dm6 coordinates.

Strain / species	(CA)n units	H-x-H	Note
Within *D. melanogaster* (strains)			
D. melanogaster dm6 (ISO1, ref.)	9	Y	reference assembly
OreR / w1118 (lab)	9	Y	identical to reference
Denmark (DK)	6	Y	6-bp in-frame deletion (tract shortened)
Spain (ES)	8	Y‡	2- & 1-bp indels (frameshift)
Finland (FI)	6	Y	6-bp in-frame deletion (tract shortened)
Sweden (SE)	10	Y‡	2-bp deletion (frameshift)
Turkey (TR)	8	Y‡	2- & 1-bp indels (frameshift)
Ukraine (UA)	9	Y	identical to reference
Zambia (ZI)	8	Y‡	2- & 1-bp indels (frameshift)
Across *Drosophila* (species)			
D. melanogaster	9	Y	~0 Mya; reference
D. simulans	6	Y	~3-5 Mya; strong
D. sechellia	7	Y	~3-5 Mya; strong
D. mauritiana	9	Y	~3-5 Mya; strong
D. yakuba	7	Y	~10-13 Mya; strong
D. erecta	5	Y	~10-13 Mya; strong
D. ananassae	0	N	~15-20 Mya; moderate; tract collapsed
D. pseudoobscura	5	N	~25 Mya; partial; ORF truncated pre-tract
D. virilis	0	N	~40-50 Mya; fragmentary; ORF truncated pre-tract

Table S2. (CA)n tract length and His-motif status at the ZMEG locus within *D. melanogaster* and across *Drosophila*. *‡, His motif present but the (CA)n tract is frameshifted (in-frame His run disrupted).*

'DataS1.csv'

Data S1. Per-protein H-x-H / C-x-x-C cluster counts and codon usage for neORFs, the *Drosophila melanogaster* proteome, and random ORFs.