

Retrotransposed gene copies persist under relaxed selection in wild *Spodoptera frugiperda*

Kiwoong Nam ^{1*}, Sylvie Gimenez ¹, Hyerin An ¹, Karine Durand ¹, Melanie Gasser ¹,
Sudeeptha Yainna ¹, Fabrice Legeai ², Julien Beuzelin ³, David G. Heckel ⁴, Sabine Hänniger ^{4,5},
Emmanuelle d'Alençon ¹

¹DGIMI, INRAE, Univ Montpellier, Montpellier, France

²INRAE, UMR-IGEPP, Bioinformatics Platform for Agroecosystems Arthropods, Campus Beaulieu, Rennes 35042, France

³Everglades Research & Education Center, University of Florida Institute of Food and Agricultural Sciences, Belle Glade, FL 33430, USA

⁴Entomology Department, Max Planck Institute for Chemical Ecology, Jena, Germany

⁵Institute for Biodiversity, Ecology and Evolution, Friedrich Schiller University Jena, Jena, Germany

*Corresponding author: Email: ki-woong.nam@inrae.fr.

Associate editor: Russell Corbett-Detig

Abstract

Copy number variations (CNVs) are major sources of genetic variation, affecting large portions of genomes. Typical CNVs are known to have short evolutionary half-lives due to deleterious effects, with documented events of adaptive evolution. However, it is not known whether the same trends apply to CNVs preserving intact Open Reading Frames (iORF-CNVs), which can encode functional proteins. Here, we investigate the mechanistic origins and evolutionary dynamics of iORF-CNVs in *Spodoptera frugiperda*, a globally distributed pest species where CNV-mediated adaptive evolution has been reported. Using PacBio HiFi reads, we generated high-quality, individually assembled genomes for 36 field-collected individuals and identified 349 genes with polymorphic iORF-CNVs. These elements are short, intronless, and display molecular signatures of LINE-mediated retrotransposition, including *cis*- and *trans*-acting effects. A majority of iORF-CNVs (68.0% to 77.28%) exhibited at least one molecular signature of retrotransposition. iORF-CNVs showed lower divergence at the first and second codon positions than at the third codon positions, but they had higher nonsynonymous-to-synonymous polymorphism ratios than non-iORF-CNV genes, implying purifying selection with relaxed constraint. Orthology analyses indicated that these CNVs originate from genes under weak evolutionary constraint, while transcriptomic and promoter motif analyses revealed that many are transcribed and retain regulatory features. We conclude that observed iORF-CNVs are generated through retrotransposition from weakly constrained genes and that the duplicates are selectively maintained for coding potential, albeit under reduced purifying selection compared with non-iORF-CNV genes. These results imply that pseudogenization is not an inevitable evolutionary fate of retrotransposed gene copies when the open reading frame is preserved.

Keywords copy number variation, individually assembled genome, intact ORF, long interspersed nuclear element, retrotransposition, *Spodoptera frugiperda*

Introduction

Copy number variations (CNVs) are major sources of genetic variation. In humans, for example, CNVs affect nearly 30% of the genome (Zhang et al. 2009), while single nucleotide variations (SNVs) comprise 84.7 Mb (Auton et al. 2015), corresponding to only 2.73% of the 3.1 Gb human genome. CNVs have been proposed as substrates for adaptive evolution through the beneficial effects of changes in gene expression resulting from different copy numbers (Spealman et al. 2025). Duplicated genes can contribute to evolutionary innovations by relaxing the evolutionary constraints on the original copies (Ohno 1970; Hughes 2002; Innan and Kondrashov 2010) or by generating diverse substrates

that can be selectively targeted (Hoile et al. 2025), for example, for the immune system (Maizels 2005) or for detoxification (Calla et al. 2017). The deleterious effects of CNVs have also been extensively documented, particularly in the context of human disease (Girirajan et al. 2011; Pös et al. 2021).

The evolutionary forces acting on the individual duplicated gene copies present within CNVs have been extensively studied over the past two decades, revealing a general trend. Population genetic analyses of polymorphic CNVs suggest that duplicated copies typically have much shorter evolutionary half-lives than SNVs, likely due to the deleterious effects (Redon et al. 2006; Emerson et al. 2008; Sudmant et al. 2010; Schrider et al. 2011, 2013; Zarrei et al. 2015; Zhang et al. 2021; Zhang and

Received: August 21, 2025. Revised: April 24, 2026. Accepted: May 14, 2026

© The Author(s) 2026. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Tautz 2022; Fang and Edwards 2024), while effectively neutral CNVs may also constitute a fraction of the standing variation. There are also well-documented instances of adaptive evolution involving CNVs (Lye and Purugganan 2019; Sirén et al. 2021; Li et al. 2023; Jayakodi et al. 2024), but, given that the majority of CNVs involve truncated genes, it is perhaps unsurprising that most CNVs exhibit non-adaptive consequences.

However, it is not clear if generalization can be extended to the CNV copies with intact Open Reading Frame (iORF-CNV), which can encode functional protein-coding sequences (Fig. 1a). iORF-CNVs caused by retrotransposition have been reported in certain species (Gao et al. 2019; Nguyen et al. 2024). If retrotransposition is a general mechanism of iORF-CNVs and iORF-CNVs generally lack promoter sequences during retrotransposition, iORF-CNVs would not be functionally constrained and would eventually become pseudogenes. Alternatively, if iORF-CNVs are transcribed but exert deleterious effects by altering gene expression levels, pseudogenization might also be favored through positive selection of nonsense or frameshift mutations. Conversely, if changes in protein sequences are more deleterious than changes in gene expression, iORF-CNVs may be subject to functional constraint. Importantly, we cannot exclude the possibility that the observed iORF-CNVs are generally targets of positive selection.

Retro-CNVs are believed to be caused by transposable elements (TEs) through retrotransposition of cytoplasmic processed transcripts. In particular, Long Interspersed Nuclear Elements (LINEs) exert *trans*-effects of retrotransposition when the reverse transcriptase recognizes the poly-A tails of cytoplasmic transcripts (Doucet et al. 2015), leading to the retro-insertion of these transcripts into the genome (Esnault et al. 2000). Such retrotransposed copies exhibit distinct footprints, including being intronless and having poly-A tails with target site duplications at both ends of the inserted gene. LINEs also exhibit *cis*-effects, where genes located near LINEs are co-transcribed in a single transcript and co-inserted into the genome (Moran et al. 1999). This LINE-mediated transduction has been frequently observed (Goodier et al. 2000; Kim et al. 2019; Schloissnig et al. 2025). Theoretically, iORF-CNVs can be generated through retrotransposition either by the *trans*- or *cis*-effects of LINEs. Alternatively, iORF-CNVs may also arise through non-retrotransposition processes, such as non-allelic homologous recombination, non-homologous end joining during DNA repair, or errors occurring during DNA replication (Hastings et al. 2009), particularly if most CNVs are produced by non-allelic homologous recombination (Schridder and Hahn 2010).

To address the formation mechanisms and evolution of iORF-CNVs and the evolution of iORF-CNVs, we focus on the fall armyworm (*Spodoptera frugiperda*; Lepidoptera; Insecta), a highly polyphagous pest insect species that feeds on more than 353 plant species across 76 families (Montezano et al. 2018). It comprises two host-associated strains, the corn strain and the rice strain, named after their preferred host plants in their native range across the Americas (Pashley 1986). Since its initial report in West Africa in 2016 (Goergen et al. 2016), invasive *S. frugiperda* has rapidly expanded its range across Africa, Asia, Oceania, and, most recently, Europe. Invasive *S. frugiperda* was identified as belonging to the corn strain (Yainna et al. 2022; Durand et al. 2024), thereby causing severe losses in maize production, for example, reducing yields by between 21% and 53% in sub-Saharan Africa (Day et al. 2017). Hence, the spread of

S. frugiperda is exerting significant pressure on global agriculture (Overton et al. 2021).

In *S. frugiperda*, CNVs are overrepresented in genes involved in interactions with host plants, particularly detoxification and digestion genes (Gouin et al. 2017). Compared to monophagous lepidopteran species, *S. frugiperda* harbors an increased copy number of detoxification genes followed by diversification (Gouin et al. 2017). This diversification may have contributed to the ability to metabolize a broad range of plant defense compounds from diverse plant species. Additionally, positive selection for increased copy numbers of insecticide resistance genes has been reported in populations from Puerto Rico (Gimenez et al. 2020), where insecticides are intensively used for maize seed production (Belay et al. 2012).

In this study, we investigated the mechanisms and evolutionary patterns of iORF-CNV polymorphisms in *S. frugiperda* using high-quality, individually assembled genomes from a natural population. These assemblies provide a maximally complete representation of each organism's genetic makeup, capturing structural variations, including CNVs. Moreover, these assemblies enable detailed sequence analyses for each individual, allowing the identification of polymorphic patterns that are potentially linked to the mechanistic origins and evolutionary forces underlying iORF-CNVs.

Our research was structured into three primary objectives. First, we generated individual genome assemblies from 36 samples to create a dataset that robustly represents the population, followed by the identification of genes showing iORF-CNVs. Second, we investigated the molecular mechanisms responsible for the generation of these iORF-CNVs. Third, we investigated the evolutionary fate of the identified iORF-CNVs.

Results

Sequencing and genome assembling

We first established a bioinformatics pipeline to generate high-quality individual genome assemblies using one-sixth of a PacBio HiFi flow cell, optimizing the approach with a laboratory colony before applying it to a large number of samples from a natural population. High-molecular-weight gDNA was extracted from a single insect of a laboratory *S. frugiperda* strain (Gouin et al. 2017) and PacBio HiFi sequencing was performed, yielding 28 Gb from a single flow cell. The one-sixth subset of the flow cell was 4.6 Gb, which corresponds to 11.5X coverage for the 400 Mb genome (Gouin et al. 2017; Durand and Nam 2025). Among three assemblers, including wtdbg2 (Ruan and Li 2020), raven (Vaser and Šikić 2021), and hifiasm (Cheng et al. 2021), wtdbg2 produced the highest BUSCO scores and contiguity (Table S1). Consequently, we selected wtdbg2 as the assembler. When the resulting assembly was polished three times with racon (Vaser et al. 2017), the number of “Single and Complete” BUSCOs increased from 5,134 to 5,148, a value even higher than that of the ver7 reference genome assembly (5,103) (Fiteni et al. 2022). Therefore, we established a bioinformatics pipeline consisting of wtdbg2 followed by three rounds of racon polishing for genome assembly.

Subsequently, high-molecular-weight gDNA was extracted from 52 fresh insect samples collected in a maize field in Florida (see “Material and methods” for details). In total, 36 samples yielded >4 µg of non-degraded gDNA (Figure S1). PacBio

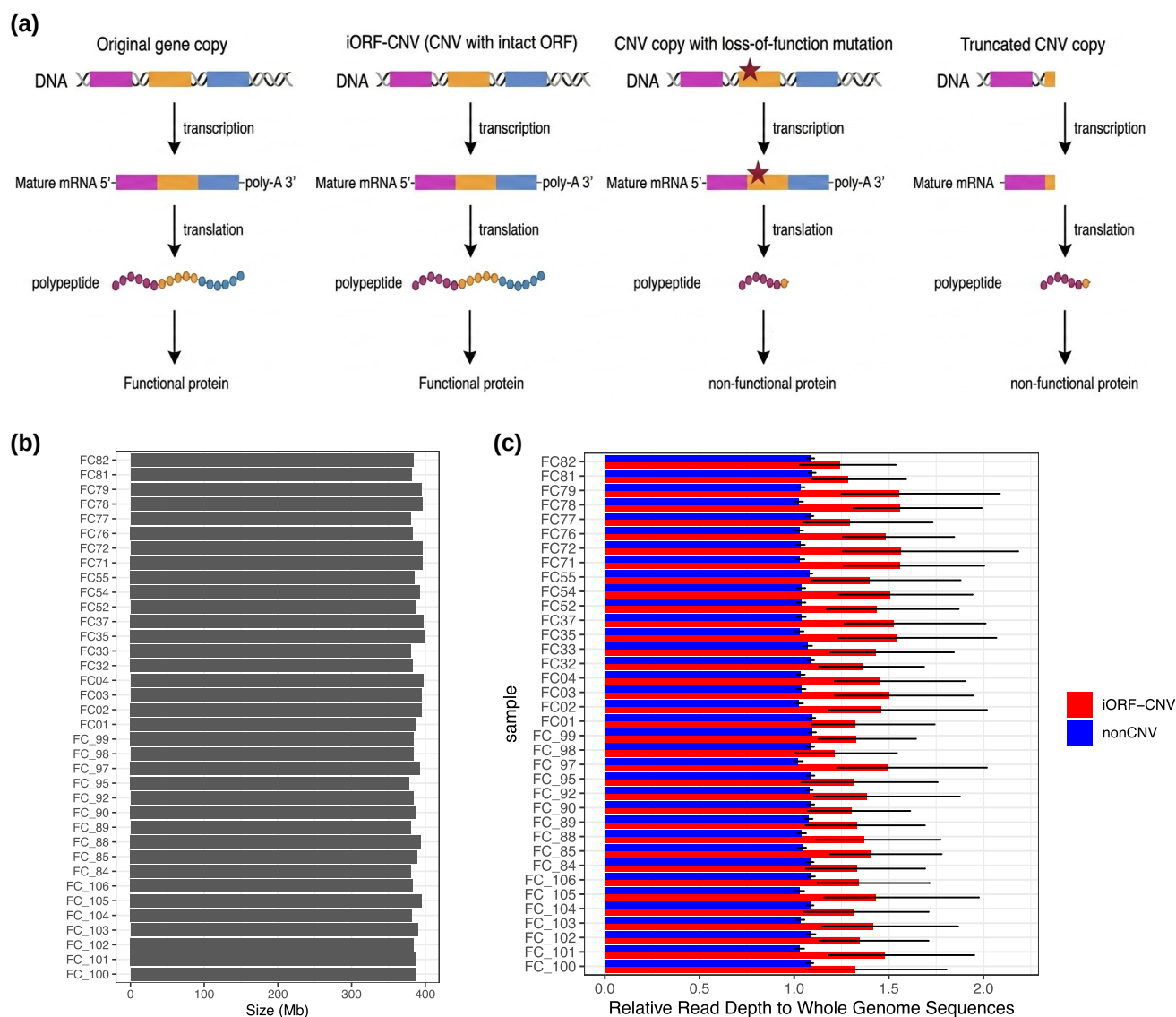


Figure 1 a) Schematic representation of iORF-CNVs. If a duplicated gene copy maintains an intact open reading frame (iORF-CNV), a functional protein can be generated. Conversely, nonfunctional proteins are produced if the duplicated copy contains a loss-of-function mutation (indicated by a red star), such as a nonsense or frameshift mutation, or if the duplicated copy is truncated. b) The size of individually assembled genomes. c) The read depth of iORF-CNVs and non-iORF-CNV genes normalized by the genomic read depth.

HiFi sequencing was performed on these 36 samples using six flow cells, with six samples multiplexed per flow cell. The average sequencing throughput per sample was 4.61 Gbp (11.53X coverage, assuming a genome size of 400 mb), with a range from 3.12 Gbp (7.8X) to 10.39 Gbp (25.98X) (Figure S2). The size of the individually assembled genome sequences ranged from 378.1 to 398.8 Mb (Fig. 1b), with N50 values ranging from 352 Kb to 3.710 Mb, and an average of 1.168 Mb (Figure S3). These assembly sizes closely matched the genome size measured by flow cytometry, which was 396 ± 3 Mb (Gouin et al. 2017). Assembly sizes exhibited a bimodal distribution according to sex (ZZ in males and ZW in females in Lepidoptera), suggesting that the presence of the W chromosome increased the assembly size in females (Figure S4).

The proportion of Complete and Single-copy BUSCO genes was consistently higher than 95% in all samples (95.8% to 97.6%), except for one sample (FC72, 93.9%), which was comparable to the

reference genome assembly (96.5%) (Figure S5). The proportion of Complete and Duplicated BUSCO genes ranged between 0.4% and 1.4%, which was also comparable to the reference genome assembly (1.0%) (Figure S5). Assuming that highly similar sequences within a single assembly can result from artificial duplications, we searched for sequence pairs longer than 100 kb with over 98% similarity through self-alignment. No assembly exhibited such cases, indicating that there were no obvious artificial duplications. Furthermore, the mappings of each assembly to the reference genome assembly show clear collinearity (Figure S6).

We annotated genes for each genome assembly by mapping 14,679 unique protein sequences from the RefSeq dataset, which was originally derived from the repeat-masked reference genome. The proportion of mapped genes ranged from 99.30% to 99.54% of the total RefSeq genes among the 36 assemblies. We filtered out those with protein sequence identities below 95% or coverage below 95% and discarded mappings containing internal stop codons to

ensure the identification of genes with intact ORFs. After this filtering, the remaining genes ranged from 7,973 to 10,227, corresponding to 54.11% to 68.68% of the total RefSeq genes (Figure S7). Since the varying number of genes might affect downstream analyses, we analyzed only the genes identified in at least two-thirds of the total samples (ie 24 samples), resulting in 7,945 genes.

Identification of iORF-CNVs

To identify polymorphic iORF-CNVs, we analyzed variation in the number of mappings among genome assemblies. Since misassemblies could affect the number of mappings, we first analyzed the mapping of lepidopteran BUSCO genes, assuming that each should map only once per genome. Although BUSCO genes are not necessarily absolute single-copy genes (Simão et al. 2015), multiple mappings were assumed to be potential indicators of misassembly. The vast majority of BUSCO genes (98.85% of 5,224 genes) had duplications from fewer than four samples (Figure S8). Thus, multiple mappings observed in at least four independent genome assemblies were considered unlikely to result from misassemblies.

Multiple mappings from at least four samples were observed in 371 RefSeq genes. To exclude TEs from candidate iORF-CNVs, we removed genes where 90% of the sequence was TE-derived and genes with TE-like annotations. After this filtering, 349 RefSeq genes were retained, and we considered that these genes exhibited iORF-CNVs in the population (see Table S2 for the full list). These 349 genes were clustered into 288 distinct groups based on the same sequence identity and coverage criteria used for mapping (95% sequence identity and 95% coverage).

To further validate the presence of polymorphic iORF-CNVs, we calculated the read depth of the mappings against the reference genome. The polymorphic iORF-CNVs consistently exhibited higher read depth than the other genes by 11.8% to 52.3% (FDR corrected P -values < 0.005 , permutation test with 1,000 replications), supporting the increased copy numbers of the identified iORF-CNVs (Fig. 1c). The RefSeq genes showing iORF-CNVs spanned a total of 2,647,800 bp, accounting for 0.642% of the reference genome (Figure S9). This length is substantially lower than the total length of SNVs identified from the HiFi reads (24,714,105 bp), implying that the genomic proportion occupied by iORF-CNVs is smaller than that affected by SNVs.

The iORF-CNVs included 12 genes encoding cuticle proteins, 12 chorion genes, and 5 trypsin genes (Table S3), along with 196 genes of unknown function. In total, 17 Gene Ontology terms were overrepresented in the list of polymorphic iORF-CNVs (Table S4), including three chorion-related ones, such as “chorion-containing eggshell formation” (GO:0007304, odd ratio = 322.49, FDR-adjusted P -value = 3.32×10^{-14}). A cuticle-related Gene Ontology term, “structural constituent of cuticle” (GO: 0042302, odds ratio = 4.22, FDR-adjusted P -value = 3.02×10^{-3}), was also overrepresented. This result indicates a nonrandom association between the iORF-CNVs and their functions.

Retrotransposition as a cause of iORF-CNVs

The average length of the iORF-CNVs in each assembly ranged from 1,352 to 1,637 bp, substantially shorter than that of

non-iORF-CNVs, which ranged from 6,654 to 8,212 bp, by a factor of 4.43 to 5.65 (FDR-adjusted P -values $< 2.2 \times 10^{-16}$ for each genome assembly, two-tailed Wilcoxon signed-rank test) (Fig. 2). Similarly, iORF-CNVs had fewer introns (range: 0.515 to 0.708) than non-iORF-CNV genes (range: 3.461 to 3.698), by a factor of 5.03 to 7.05 (FDR-adjusted P -values $< 2.2 \times 10^{-16}$). The proportion of intronless ORF-CNVs (50.79% to 67.46%) was significantly higher than that of non-iORF-CNV genes (15.45% to 16.78%, FDR-adjusted P -values $< 2.2 \times 10^{-16}$, two-tailed Fisher's exact test). To control for the effect of gene length, we analyzed the proportion of intronless genes within specific size bins. In the 0 to 1,000 bp range, the proportion of intronless ORF-CNVs (81.98% to 89.12%) remained significantly higher than that of non-iORF-CNV genes (49.66% to 65.32%) (FDR-adjusted P -values $< 2.2 \times 10^{-16}$). Similarly, within the 1,000 to 2,000 bp range, the intronless proportion of ORF-CNVs (52.14% to 62.88%) was also significantly greater than that of non-iORF-CNV genes (26.69% to 31.49%) (FDR-adjusted P -values $< 1.99 \times 10^{-13}$). Among the intronless genes, iORF-CNVs were consistently shorter than non-iORF-CNV genes across all samples with statistically significant differences (FDR-adjusted P -values < 0.1), except for one sample (FDR-adjusted P -values = 0.1907). However, the magnitude of these differences was relatively small, ranging only between 0.23% and 17.18%. This result suggests that the observed difference in gene length is primarily driven by intron numbers, proposing the possibility that a substantial proportion of iORF-CNVs were generated through retrotransposition.

To investigate the genomic context of the identified polymorphic iORF-CNVs, we mapped a pair of 100 kb flanking sequences of each iORF-CNV gene from the individual assemblies back to the reference genome. These flanking sequences of all 349 iORF-CNV genes were successfully localized to specific genomic coordinates. In the vast majority of cases (91.98%; 321/349), the flanking sequences mapped to adjacent regions in the reference genome without intervening genes. This result suggests that these iORF-CNVs originated from insertion events into non-genic regions in the reference individual, although the possibility of gene excision in the reference genome cannot be entirely excluded. Among these insertions, 63.6% (222/349) shared the same flanking sites across multiple individuals, indicating that they represent shared alleles derived from common ancestral insertion events. The singleton insertions ($n = 99$) are noticeable, as these insertions indicate that single genes are inserted into different genomic sites multiple times independently.

When the genomic location of each iORF-CNV copy was examined, 75.86% to 84.41% of the polymorphic iORF-CNV copies involved interchromosomal duplications (Fig. 3a). Among them, intronless copies had lower proportions of exclusive intrachromosomal duplications (6.33% to 17.91%) than intronic copies (22.22% to 40%), with statistical significance observed in 34 out of 36 assemblies (FDR-adjusted P -values < 0.1 , permutation test with 1,000 replications) (Fig. 3b). This result indicates that the iORF-CNVs frequently involve interchromosomal duplications with a stronger tendency for intronless genes, consistent with the expectation that a substantial proportion of the iORF-CNVs originated through retrotransposition.

Then, we tested the *trans*-effects of LINES, which typically result in retrotransposed sequences flanked by poly-A tails and target site duplications (Richardson et al. 2014). To test whether the polymorphic iORF-CNVs exhibit these structures, we searched for

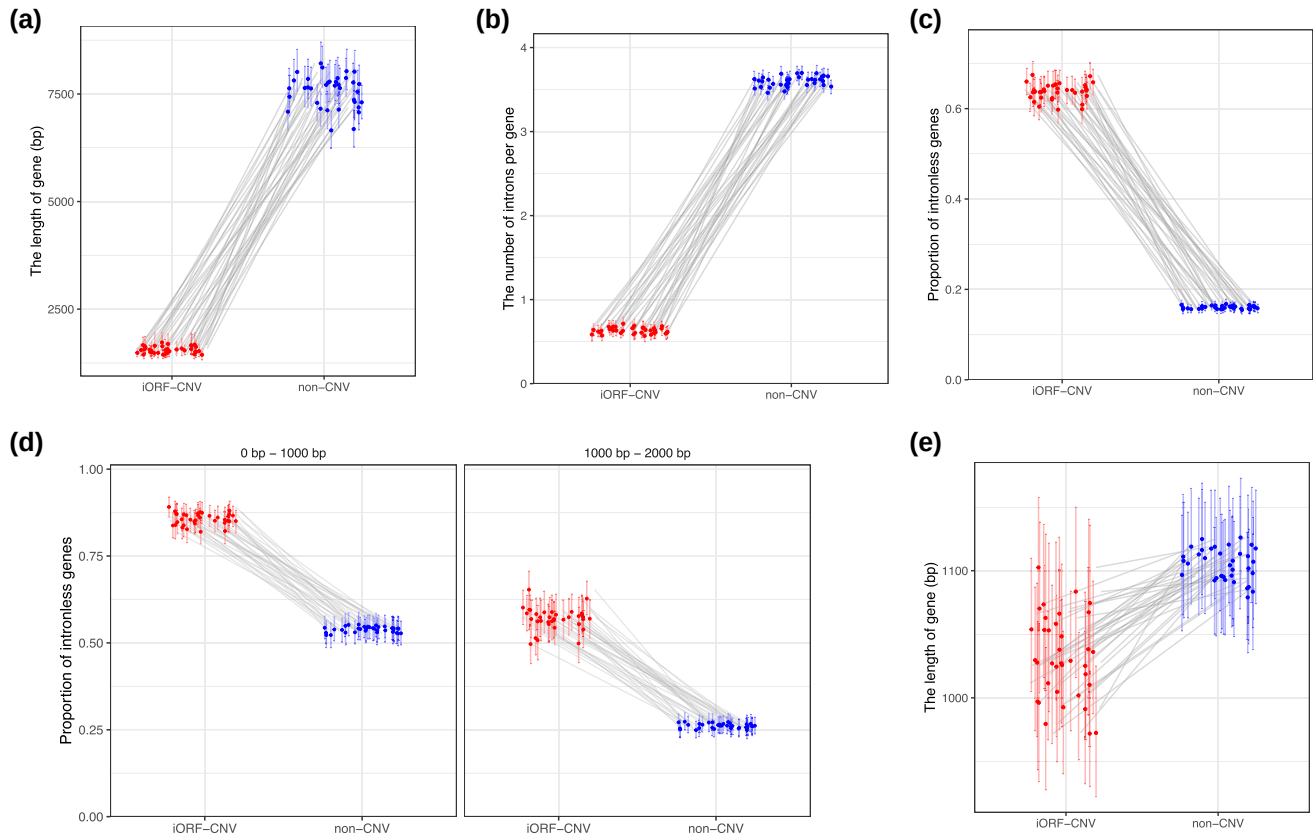


Figure 2 The structure of iORF-CNV. Comparison of a) gene length, b) number of introns, c) proportion of intronless genes, d) proportion of intronless genes within specific length bins, and e) gene length of single-exon genes between genes with multicopy iORF-CNVs (red) and non-iORF-CNV genes (blue) in a genome for each sample. Error bars represent 95% confidence intervals calculated from bootstrapping based on 1,000 resamplings of 100 kb genomic windows.

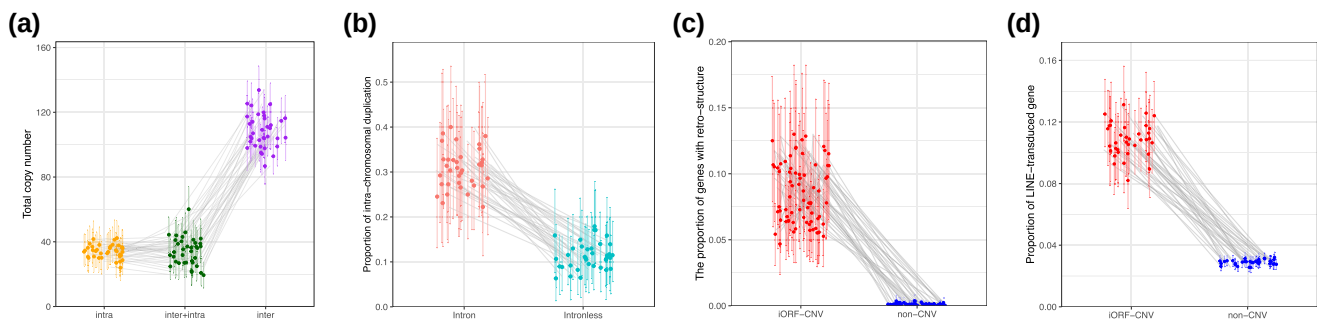


Figure 3 The possibility of retrotransposition causing polymorphic iORF-CNVs. a) Number of iORF-CNV copies classified as exclusively intrachromosomal, exclusively interchromosomal, or a combination of inter- and intrachromosomal duplications for each sample. b) Proportion of exclusively intrachromosomal duplications among intronless and intron-containing iORF-CNVs. c) Proportion of iORF-CNVs exhibiting retrotransposition-derived structural features, including the presence of target site duplications, the intact gene body, and a poly-A tail. d) Proportion of iORF-CNVs associated with LINE-mediated transduction events. Error bars represent 95% confidence intervals calculated from bootstrapping based on 1,000 resamplings of 100 kb genomic windows.

the presence of poly-A tails and target site duplications at both ends of each iORF-CNV or non-iORF-CNV gene for each genome assembly. To ensure stringency, we considered only poly-A tails that were ≥ 20 bp in length and target site duplications that were ≥ 5 bp. iORF-CNVs showed a higher proportion of these structures (4.71% to 8.25%) than non-iORF-CNV genes (0.195% to 0.273%) (FDR-corrected P -value $< 2.277 \times 10^{-18}$, two-tailed Fisher's exact test) (Fig. 3c). This result supports the hypothesis that a

substantial proportion of the iORF-CNVs originated through retrotransposition via *trans*-effects of LINE.

We also tested LINE-mediated transduction through *cis*-effects. iORF-CNV genes exhibited higher proportions of cases in which the gene is flanked by a LINE sequence (8.20% to 13.12%) than non-iORF-CNV genes (2.55% to 3.21%) by a factor of 2.77 to 4.74 (FDR-adjusted P -values $< 1.7 \times 10^{-11}$, Fig. 3d). This result supports the role of the *cis*-effect of LINE through transduction.

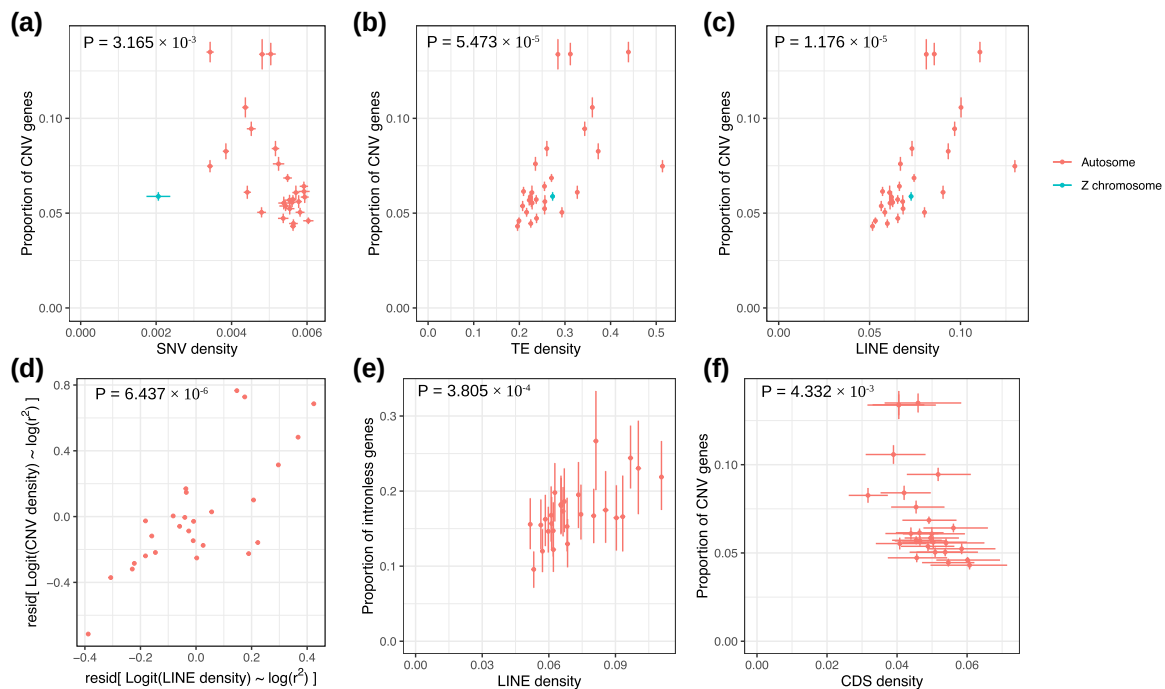


Figure 4 The distribution of iORF-CNVs and transposable elements. Correlation of iORF-CNV density calculated for each chromosome with a) SNV density, b) transposable element density, and c) LINE density. d) Partial correlation analysis between LINE density and iORF-CNV density, controlling for the effect of linkage disequilibrium (estimated by r^2). e) Correlation between LINE density and the proportion of intronless genes. f) Correlation between protein-coding sequence density and iORF-CNV density. Error bars represent standard errors, calculated from sample-level statistics across genomes.

To quantitatively estimate the extent to which retrotransposition contributes to the genesis of iORF-CNVs, we calculated the proportion of iORF-CNVs satisfying three distinct retrotransposition hallmarks across all 36 genome assemblies, including intronless, interchromosomal duplication, and LINE-mediated insertion footprints. A substantial majority of iORF-CNVs, ranging from 68.0% to 77.28%, exhibited at least one of these indicators (Figure S10). Notably, a significant proportion (35.68% to 43.44%) satisfied at least two criteria, while 8.75% to 12.80% met all three conditions concurrently. These results suggest that retrotransposition is one of the major drivers of iORF-CNV formation.

Shared evolutionary forces determining iORF-CNV and TE element distribution

Then, we inferred the evolutionary forces determining the spatial distribution of iORF-CNVs along the genome. The iORF-CNV density calculated for each chromosome was negatively correlated with SNV density (Spearman's $\rho = -0.535$, P -value = 0.003165) (Fig. 4a), suggesting that the distribution of CNVs is shaped by evolutionary forces different from those affecting point mutations. In contrast, iORF-CNV density exhibited a positive correlation with both total TE ($\rho = 0.30$, P -value = 5.473×10^{-5}) and LINE densities ($\rho = 0.732$, P -value = 1.176×10^{-5}) (Fig. 4b and c). This result suggests that the distributions of TEs and polymorphic iORF-CNVs can be influenced by the same evolutionary forces, such as shared insertional preferences.

We indirectly tested this possibility using an autosomal correlation test between iORF-CNV and LINE densities while controlling for the effect of purifying selection. We assumed that linkage

disequilibrium can be used as a proxy for the inefficacy of selection, as linked selection reduces local effective population size (Smith and Haigh 1974; Charlesworth 1996). The SNV density showed a negative correlation with the r^2 of allele frequencies, used as a measure of linkage disequilibrium strength ($\rho = -0.5360$, P -value = 0.003133), suggesting that this assumption is valid. iORF-CNV density was positively correlated with r^2 ($\rho = 0.5517$, P -value = 0.00225). After accounting for the effects of r^2 on both iORF-CNV density and LINE density using partial regression, the positive correlation between LINE density and iORF-CNV density remained (P -value = 6.437×10^{-6} , Fig. 4d), supporting the possibility of a shared insertional preference between LINE and iORF-CNVs.

Shared insertional preferences between LINEs and iORF-CNVs were further tested with a focus on intronless genes. We assumed that most intronless genes experienced retrotransposition over long evolutionary periods (Grzybowski 2012), during which detectable LINE insertions occurred in the genome. The LINE density was positively correlated with the proportion of intronless genes in the reference genome ($\rho = 0.6453$, P -value = 0.0003805, Fig. 4e), supporting this possibility. Coding sequence density was negatively correlated with iORF-CNV density ($\rho = -0.6404$, P -value = 0.0004332, Fig. 4f), suggesting that regions with lower coding density offer higher opportunities for both LINE and iORF-CNV insertion than regions with lower density.

Selective constraints on iORF-CNVs

To test the evolutionary constraints on the iORF-CNVs, we calculated the pairwise divergence of coding sequences among iORF-CNV copies originating from the same genes within a single

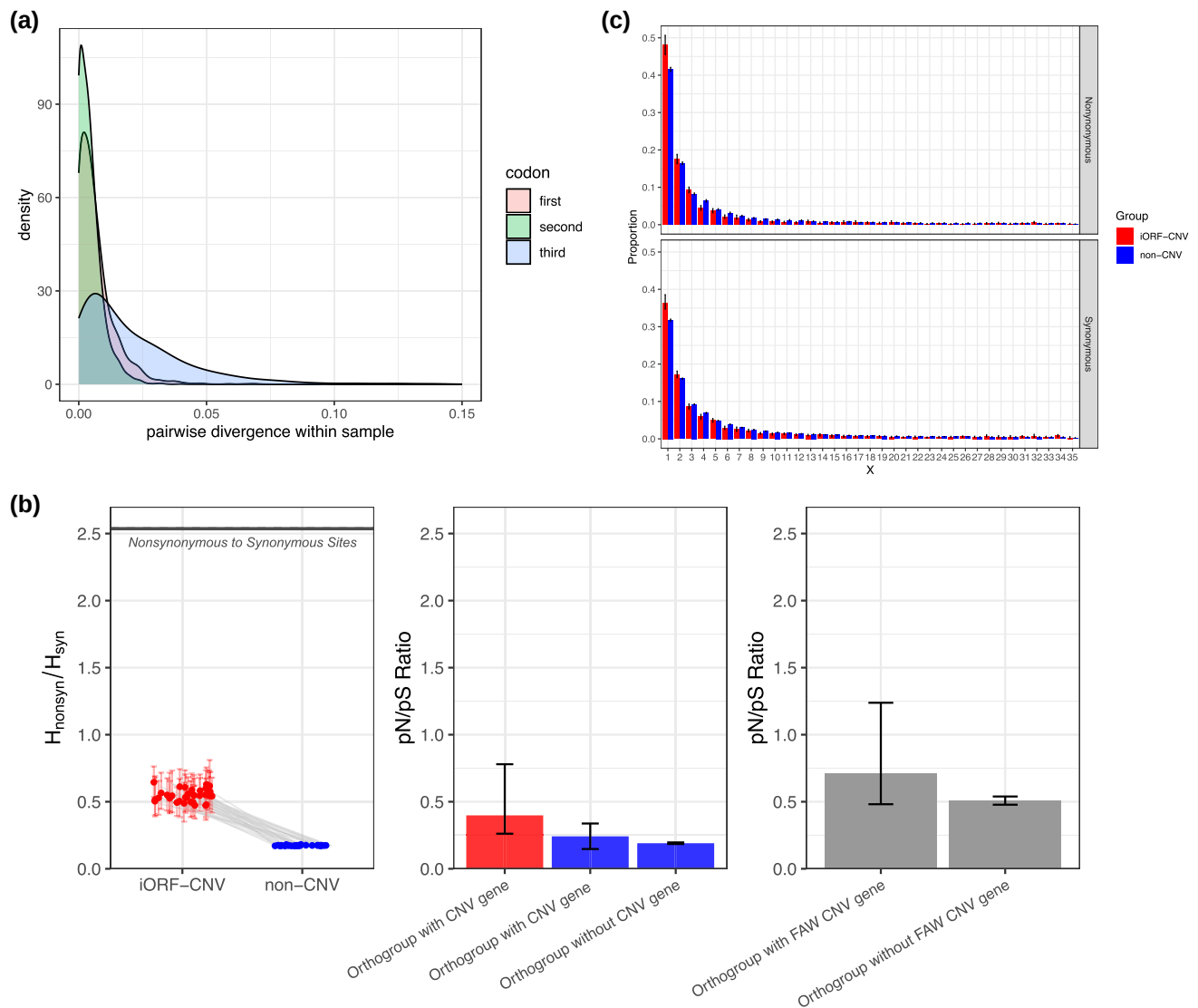


Figure 5 Evolutionary forces underlying polymorphic iORF-CNVs. a) Density plot of pairwise nucleotide divergence among iORF-CNV copies within a single genome, calculated separately for the first, second, and third codon positions. b) (Left) Ratio of nonsynonymous missense to synonymous heterozygous positions for each sample in iORF-CNV genes versus non-iORF-CNV genes. Horizontal lines indicate the ratio of nonsynonymous missense to synonymous sites in the same gene sets. (Middle) Ratio of nonsynonymous missense to synonymous polymorphisms in iORF-CNV genes, non-iORF-CNV genes within orthogroups containing iORF-CNVs, and genes from orthogroups without iORF-CNVs. (Right) Ratio of nonsynonymous missense to synonymous polymorphisms in *S. exigua*, comparing orthologous genes with or without iORF-CNVs in *S. frugiperda*. c) Folded site frequency spectra of nonsynonymous and synonymous polymorphisms in iORF-CNV and non-iORF-CNV genes. Error bars represent 95% confidence intervals calculated from bootstrapping based on resamplings of genes.

assembly. Third codon positions exhibited higher divergences (2.371%, 95% CI: 2.271% to 2.471%) than first (0.686%, 95% CI: 0.665% to 0.706%) or second codon positions (0.429%, 95% CI: 0.418% to 0.441%) (Fig. 5a), which is expected for evolutionarily conserved coding genes. iORF-CNVs displayed a higher pN/pS (nonsynonymous to synonymous polymorphisms ratio) (0.4722 to 0.6454) than non-iORF-CNV genes (0.1708 to 0.1813, FDR-corrected P -value $< 2.2 \times 10^{-16}$; Fisher's exact test) (Fig. 5b). However, these pN/pS values are much lower than the ratio of nonsynonymous to synonymous sites (2.536, 95% CI: 2.534 to 2.539), implying underrepresentation of nonsynonymous variations at iORF-CNVs. These results suggest that polymorphic iORF-CNVs are under selective constraints, but to a lesser extent than non-iORF-CNV genes.

The lower pN/pS observed in iORF-CNVs than in non-iORF-CNV genes may result from relaxed selective constraints following duplication. Alternatively, and not mutually exclusively, this pattern could indicate that the iORF-CNVs originated from genes that already had lower selective constraints before duplication than the rest of the genome. To test these possibilities, pN/pS was calculated from the iORF-CNVs and non-iORF-CNVs genes within the same orthogroups. iORF-CNVs exhibited a higher pN/pS ratio (0.3983; 95% CI: 0.2616 to 0.7797) than non-iORF-CNV genes within the same orthogroups (0.2418; 95% CI: 0.1474 to 0.3375; P -value $< 2.2 \times 10^{-16}$; odds ratio = 1.647) (Fig. 5c). These non-iORF-CNV genes, in turn, had higher pN/pS than the genes in orthogroups without iORF-CNVs (0.1904; 95% CI: 0.1847 to 0.1967; P -value $< 2.2 \times 10^{-16}$; odds ratio = 1.270). This result

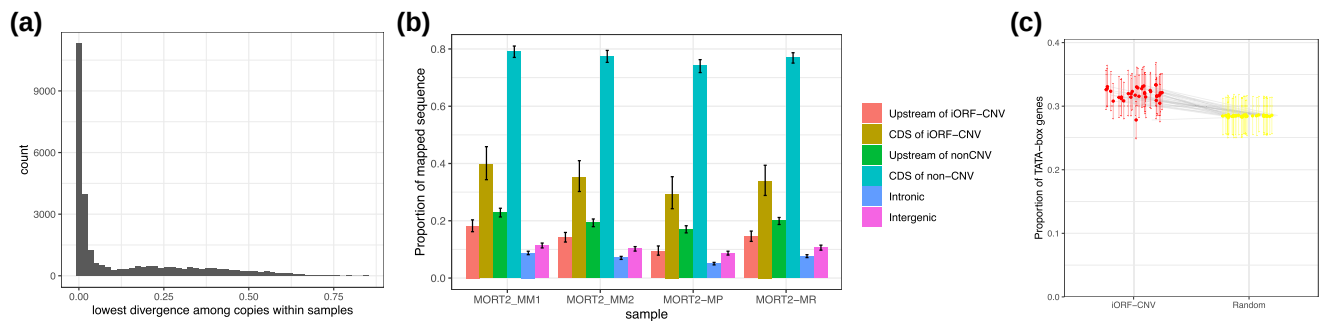


Figure 6 Transcription of iORF-CNVs. a) Distribution of the lowest pairwise divergence among 300 bp upstream sequences of iORF-CNV copies within a genome. b) Proportion of sequences with mapped reads in 300 bp upstream regions of iORF-CNV and non-iORF-CNV genes, as well as intronic, intergenic, and protein-coding sequences. c) Proportion of 300 bp upstream regions containing TATA-box motifs for iORF-CNV (red) and 1,000 randomly selected genomic sequences (yellow) for each sample. Error bars represent 95% confidence intervals calculated from bootstrapping based on resamplings of b) 100 kb windows or c) genes.

supports the hypothesis that iORF-CNVs originated from genes already under lower constraints and, also, that the selective constraint was further reduced after duplication.

To test this hypothesis further, we analyzed resequencing data from 16 samples of *S. exigua*. Orthogroups containing *S. frugiperda* iORF-CNV genes exhibited higher pN/pS in *S. exigua* (0.7152, 0.4815 to 1.2381 of 95% CI) than orthogroups without *S. frugiperda* iORF-CNV genes (0.5088, 0.4777 to 0.5391 of 95% CI) (P -value = 0.005463, odds ratio = 1.406) (Fig. 5c). This result also support that iORF-CNVs originate from genes already under weak selective constraint prior to duplication.

The increased pN/pS in iORF-CNVs could also be explained if a significant proportion of CNV copies are experiencing positive selection through the increase in the frequency of nonsynonymous alleles at the time of sample collection. If true, we expect iORF-CNVs to have a higher proportion of derived alleles for nonsynonymous polymorphisms than non-iORF-CNV genes. However, we observed no clear difference in the folded site spectrum between iORF-CNVs and non-iORF-CNV genes, except for singleton polymorphisms (Fig. 5d), which should be primarily determined by demographic events (Cubry et al. 2017). This result suggests that positive selection is unlikely to be the main evolutionary force driving the increased pN/pS in iORF-CNVs.

Promoter sequences of iORF-CNVs

Because evolutionary constraint on iORF-CNVs is expected only if they are transcribed, we tested the presence of potential transcriptional regulatory elements. We focused on the 300 bp upstream flanking sequences, which are typically considered the minimum length required to identify promoter regions in eukaryotes (Lai et al. 2019). A total of 60.32% of copies within iORF-CNVs exhibited 300 bp of upstream flanking sequences with <5% sequence divergence from the flanking sequences of other copies within a genome (Fig. 6a), suggesting the preservation of regulatory elements in these upstream regions, which might be involved in transcription.

Next, we tested whether iORF-CNVs and their 300 bp upstream sequences are indeed transcribed by analyzing transcriptome data from four larvae. Coding sequences of non-iORF-CNV genes showed a higher proportion of sequences with mapped RNA-seq reads (74.13% to 79.16%) than iORF-CNVs (29.36% to 39.71%)

(Fig. 6b). Both were higher than the proportions observed for intronic (5.03% to 8.75%) and intergenic regions (8.69% to 11.38%), suggesting that a subset of iORF-CNVs are indeed transcribed. The 300 bp upstream sequences also had a higher proportion of sequences with mapped reads than intronic and intergenic sequences for both iORF-CNVs (9.42% to 18.18%) and non-iORF-CNV genes (17.10% to 22.89%) for all investigated samples. This result indicates that iORF-CNVs and their upstream regions are transcribed more abundantly than intronic or intergenic regions.

To assess promoter functionality more directly, we searched for canonical TATA-box motifs (TATAWAW (Donczew and Hahn 2017)) in the 300 bp upstream regions. Non-iORF-CNV genes had a higher proportion of genes with TATA-box motifs than randomly chosen 1,000 positions, by 21.03% to 29.50% (FDR-corrected P -value < 2.2×10^{-16}) (Figure S11). This result suggests that TATA-box motifs can be used as indicators of promoters, even though a significant proportion of promoters are not expected to contain a TATA-box motif. The 300 bp upstream sequences of iORF-CNVs had higher proportions of TATA-box motifs (30.46% to 33.37%) than random sequences (28.17% to 28.65%, Fig. 6c) with statistical support from all samples (FDR-corrected P -values < 0.1), with a single exception (27.83%, FDR-corrected P -value = 0.627). This result supports that the upstream sequences of iORF-CNVs play a role in promoter activity.

Fixation of retrotransposed genes

Next, we investigated the evolutionary fate of iORF-CNV copies. Because we have excluded rarely occurring iORF-CNVs to minimize false positives, it was not possible to directly estimate the fixation probabilities. Instead, we focused on intronless genes, assuming that these genes experienced past events of retrotransposition. If a retrotransposed copy becomes fixed and the original is lost or pseudogenized, synteny can be disrupted among genes, particularly given the frequent interchromosomal duplications observed for iORF-CNVs (Fig. 3a and 3b).

To test this possibility, we first identified orthologous chromosomes among *S. frugiperda*, *S. litura*, and *S. exigua* using 5,286 BUSCO genes. Almost no deviations in orthologous chromosomes among these three species were observed from the BUSCO genes (Figure S12), as previously reported (Wright et al. 2024). Given that *S. frugiperda* and *S. litura* are sister taxa with *S. exigua* as an

outgroup, we assume that, if all or a subset of *S. frugiperda* genes are intronless while all genes in orthologs of the other species retain introns, this intron loss occurred specifically in the *S. frugiperda* lineage after its divergence from the lineage leading to *S. litura*. These intron-losing genes exhibited 23.4-fold and 29.7-fold greater synteny deviation than intronic and constitutively intronless genes, respectively (Figure S13, with examples), supporting the fixation of duplicated copies to a detectable extent.

Discussion

While the general mechanisms and evolutionary patterns of CNVs are well documented, the evolutionary dynamics of CNVs preserving intact Open Reading Frames (iORF-CNVs) remain insufficiently understood, leaving a gap in our comprehensive understanding of CNV evolution. To address this issue, we generated 36 individually assembled genomes from field-collected *S. frugiperda* samples using approximately 10X coverage for each sample. The sizes of all resulting assemblies are comparable to the measurements from flow cytometry, consistent with the conserved genome size within the species (Durand and Nam 2025). The BUSCO statistics reveal that these assemblies have comparable quality to the reference genome assembly, despite the relatively low coverage. We identified 349 RefSeq genes exhibiting polymorphic iORF-CNVs. The relatively higher mapping read depth of these iORF-CNVs than non-CNVs implies that the identified iORF-CNVs are robust.

These iORF-CNVs tend to be short, intronless, and display hallmarks of retrotransposition, including signatures of both *cis*- and *trans*-effects of LINES. A majority of iORF-CNVs (68.0% to 77.28%) exhibited at least one molecular signature of retrotransposition including intronless, interchromosomal duplication, and footprints of LINE-mediated insertion, showing that retrotransposition is one of the major mechanisms of the formation of iORF-CNVs. Ortholog analysis with pN/pS calculations revealed that these iORF-CNVs are evolutionarily constrained, albeit under relaxed purifying selection, and originated from genes already subject to weak selective pressure. Additionally, a substantial number of iORF-CNV copies retained upstream regulatory sequences associated with transcriptional activity, indicating a potential requirement for functional conservation. Synteny analysis further showed that a notable proportion of retrotransposed copies have been fixed in the population. Together, we conclude that a substantial proportion of iORF-CNVs originates from genes under weaker purifying selection, through retrotransposition with both *cis*- and *trans*-effects of LINES, while still being evolutionarily constrained to a lesser extent.

The evolutionary constraint observed on iORF-CNVs suggests that changes in amino acid sequences caused by mutations may be more deleterious than potential alterations in gene expression due to copy number variation. The pairwise divergence among iORF-CNV copies within a genome was 2.371% (2.271% to 2.471% of 95% CI) at third codon positions, indicating that the polymorphic copy numbers have existed long enough to generate this divergence, although the divergence estimation can be influenced by heterozygous nucleotides. Assuming a mutation rate of 2.9×10^{-9} per site per generation (Keightley et al. 2015), the average age of the iORF-CNV copies would be approximately 4 million generations ($2.371\% / (2.9 \times 10^{-9} \times 2)$), suggesting that the heterogeneous copy

numbers of iORF-CNV genes have been tolerated over this period. We postulate the possibilities that fitness effects have not been severely deleterious or that variations in copy number may not always lead to proportional changes in gene expression. Although the correlations between CNV and gene expression are well established in human cancers (Shao et al. 2019; Boen et al. 2021), such associations in Lepidoptera have been reported only for a limited number of genes (Macias-Muñoz et al. 2017; Mulhair et al. 2023). As a result, the extent to which CNVs exert genome-wide deleterious effects via altered gene expression remains uncertain. In contrast, evolutionary conservation through purifying selection by eliminating deleterious nonsynonymous mutations is a universal process across life. Thus, we propose the possibility that changes in copy numbers of the observed iORF-CNVs are more tolerable than changes in amino acid sequences.

The observation that the polymorphic iORF-CNVs originate from genes under weak purifying selection, as indicated by elevated pN/pS, suggests the possibility that genes more tolerant of nonsynonymous point mutations are also more permissive to changes in copy number. Hence, most of the polymorphic iORF-CNVs could be explained by non-adaptive processes. However, it should be noted that we do not exclude the existence of adaptive iORF-CNVs. If directional selection is involved, changes in copy numbers would rapidly be fixed in populations, and we would not see iORF-CNVs as polymorphic within a population. In the case of balancing selection involving alternative protein sequences, an elevation in pN/pS is also possible. It is not possible to estimate the selective coefficients of iORF-CNVs using population genetics analysis to evaluate their fitness effects (Eyre-Walker et al. 2006) because the individually assembled genome sequences were haplotype-collapsed. Estimating the adaptive evolutionary rates involving iORF-CNVs remains a key task and it should certainly be carried out in future studies.

Nonetheless, the nonrandom functional enrichment among iORF-CNVs, particularly in genes related to the chorion and cuticle, suggests that some may have been subject to positive selection. For example, we identified iORF-CNVs in 12 chorion genes, which are involved in eggshell formation in the ovary (Dong et al. 2022). The changes in copy number of chorion genes was also reported in Lepidoptera with the interpretation that it may be linked to adaptation against plant defense responses or hypersensitive response-like necrosis in eggs (Dort et al. 2024). We also observed iORF-CNVs in 12 cuticle genes. It is well known that cuticle thickening is associated with insecticide resistance (Balabanidou et al. 2018), including in *S. litura* (Wang et al. 2025). In *S. frugiperda*, cuticle genes are overexpressed following insecticide treatment (Zhu et al. 2021), further raising the possibility of the adaptive role of these iORF-CNVs.

Conclusion

In summary, this study investigates the evolutionary origins, mechanisms, and functional conservation of iORF-CNVs in *S. frugiperda* using high-quality, individually assembled genomes. We observed that these CNVs can arise through retrotransposition with *cis*- and *trans*-effects of LINES. A majority of iORF-CNVs (68.0% to 77.28%) exhibited at least one molecular signature of retrotransposition, indicating that retrotransposition is a primary mechanism for the formation of iORF-CNVs. Other mechanisms may also contribute to the generation of iORF-CNVs; therefore, future studies should

investigate the comprehensive mechanisms underlying these variations. We also showed that iORF-CNVs were primarily derived from genes under weak purifying selection while being selectively constrained. Hence, the evolution of iORF-CNVs appears to be different from other CNVs, most of which have lost their function during duplication due to the disruption of the Open Reading Frames. In other words, the primary evolutionary force acting on iORF-CNVs appears to be purifying selection that conserves protein-coding sequences. In contrast, other CNV genes may be more frequently shaped by positive selection favoring the fixation of loss-of-function mutations such as nonsense or frameshift changes, leading to pseudogenization, unless the duplicated copies are eliminated from the population by purifying selection against the CNVs themselves. These non-homogeneous patterns of natural selection highlight the plastic nature of the genome, which is shaped and reshaped by a complex interplay of evolutionary forces, including variation in gene number and function, TE activity, genome structure, and demographic parameters influencing the efficacy of selection.

Materials and methods

Establishing bioinformatics pipelines to generate individually assembled genomes

To ensure a sufficient sample size for population-level analysis, we developed bioinformatics pipelines capable of generating high-quality individual genome assemblies from one-sixth of a PacBio HiFi flow cell, providing approximately 10X coverage for the 400 Mb genome. High-molecular-weight genomic DNA was extracted from one half of a single pupa reared from a laboratory colony initiated with corn strain *S. frugiperda* specimens collected in Guadeloupe in 2000 (Gouin et al. 2017) using the QIAGEN DNA Midi Kit according to the manufacturer's protocol. Libraries for PacBio HiFi sequencing were generated with a 15 kb insert size using SMRTbell Express template prep, and sequencing was conducted on the PacBio Sequel II platform.

Subsequently, one-sixth of the total throughput from a single flow cell was allocated for genome assembly using wtdbg2 (Ruan and Li 2020), raven v1.4.0 (Vaser and Šikić 2021), and hifiasm v0.16.1 (Cheng et al. 2021). To polish the assembly, HiFi reads were mapped against the assemblies using minimap2 v2.5 (Li 2018), and racon v1.4.3 (Vaser et al. 2017) was used to improve the assembly with three rounds of iterations.

The quality of the assemblies produced by each assembler was assessed using BUSCO v5.2.2 (Simão et al. 2015) with lepidoptera_odb10 to identify the best assembler and to determine the effectiveness of the polishing iterations. Based on these results, we established a bioinformatics pipeline using the parameters that yielded the highest BUSCO scores to generate high-quality individual genome assemblies from the natural insect population.

Sample collection and individual genome assembly

Larval sampling was performed by handpicking from an experimental maize field at the University of Florida Institute of Food and Agricultural Sciences Everglades Research and Education

Center in Belle Glade, Florida (USA), between October and November 2021. Only one larva was collected per maize plant to minimize kinship among samples. The larvae were shipped alive to the DGIMI laboratory in France and raised to the pupal stage.

The extraction of high-molecular-weight genomic DNA and PacBio HiFi sequencing were performed as described above. A total of 36 samples were sequenced using six flow cells, yielding an average of one-sixth of a flow cell of data per sample, consistent with the procedure established during our pipeline optimization. The *de novo* genome assembly for each sample was conducted following the bioinformatics pipeline established in the previous step, by performing the initial assembly with wtdbg2 and subsequently applying three consecutive rounds of polishing with racon. The quality of the resulting assemblies was subsequently assessed using BUSCO.

To identify potential artificial duplications, self-alignment was performed for each of the 36 assemblies using minimap2 v2-2.28 (Li 2018) with the “-x asm20” preset. To ensure the detection of large-scale duplications, the bandwidth was set to 100,000 bp using the “-r” option. From the alignment results, sequence pairs longer than 100 kb with a sequence identity exceeding 98% were identified and examined, excluding exact self-matches.

To assess genomic collinearity between the 36 genome assemblies and the ver7 reference genome (https://bipaa.genouest.org/sp/spodoptera_frugiperda/download/genome/spodoptera_frugiperda_corn/v7.0/, also available at <https://doi.org/10.5281/zenodo.20487731>) (Fiteni et al. 2022), each assembly was mapped against the reference genome using minimap2 with the “-x asm20” preset. Synteny dot plots were then generated using Paf2dotplot v1.0.1 (Jiang 2026), with filtering parameters set to exclude alignment lengths and query sequences shorter than 10 kb. Finally, GraphicsMagick v1.3.38 (GraphicsMagick Group) was used to organize the resulting plots into one layout.

Read depth was calculated in 100 kb windows from the bam files using mosdepth v0.3.11 (Pedersen and Quinlan 2018). Only reads with a mapping quality of at least 20 (-Q 20) were included in the analysis. The relative depth of Z chromosomes compared to autosomes was then used to identify the sex of each sample.

Profiling iORF-CNVs

One protein sequence was selected per gene from the *S. frugiperda* RefSeq annotation (NCBI Genome ID: GCF_023101765.2). These sequences were mapped to each genome assembly using Miniprot v0.13-r248 (Li 2023). Mappings with <95% identity or coverage were discarded, and resulting GFF files were used to extract protein and coding sequences with gffread v0.9.9 (Pertea and Pertea 2020). Any mapping with internal stop codons was also removed. Potential iORF-CNV genes were identified as those with multiple mappings from a single RefSeq protein sequence. This method is robust against the inclusion of pseudogenes or fragmented copies as candidate iORF-CNV genes, as it allows for the direct verification of open reading frame integrity in each genome assembly.

Protein sequences of BUSCO genes (arthropoda_odb10) were mapped to each assembly using Miniprot v0.13-r248 (Li 2023), and mapping counts were recorded per genome. RefSeq protein sequences were also mapped to the ver7 reference genome using the same software. A TE master library was created from the ver7

reference assembly using RepeatModeler v2.0.4 (Flynn et al. 2020). TE copies were located with the resulting library in the ver7 reference assembly using RepeatMasker v4.1.5 (Smit et al. 2013).

HiFi reads were mapped to the ver7 reference genome using minimap2 v2.19 (Li 2018), and read depth was calculated from the resulting bam files using samtools v1.17 (Li et al. 2009). The ver7 reference assembly was generated from a native *S. frugiperda* population sample in Guadeloupe; therefore, the low phylogenetic distance to the field-collected samples is expected to reduce mapping artifacts and subsequently improve the accuracy of variant calling (Lefouili and Nam 2022). Gene Ontology analysis was performed using Blast2GO v 4.1.9 (Conesa and Götz 2008).

To identify the genomic positions of the iORF-CNV insertions, we first extracted a pair of 100 kb flanking sequences for each iORF-CNV gene from each of the 36 individual assemblies. These sequences were mapped to the reference genome using minimap2 v2.19 with the -x asm20 option. An insertion site was validated only if both flanking sequences were successfully mapped to a genomic region spanning a consecutive length of more than 80 kb on the reference genome. We considered an iORF-CNV as being inserted into a non-genic region if the distance between the mapped coordinates of the two inner flanking ends was <100 bp. Shared insertional sites among multiple individuals were then determined by identifying overlapping mapping positions of these flanking regions using Bedtools v2.30.0 (Quinlan and Hall 2010).

Testing retrotransposition

Inter- or intrachromosomal duplications were inferred based on the chromosomal location of mappings. Chromosome assignments for the contigs of the individually assembled genomes were inferred using the algorithm scaffold of ragtag v2.1 (Alonge et al. 2022), using the ver7 reference genome as a reference. *Trans*-effects of LINEs were inferred from iORF-CNVs flanked by a poly-A sequence (≥ 20 bp) and target site duplications (≥ 5 bp). *Cis*-effects were inferred from LINEs located within 100 bp of iORF-CNV genes. TEs in each genome assembly were identified using RepeatMasker v4.1.5 (Smit et al. 2013) with the above-mentioned TE library.

Testing natural selection

Codon-based alignments of iORF-CNV coding sequences were generated using prank v170427 (Löytynoja 2014), and Gblocks v0.91 (Castresana 2000) was used to remove poorly aligned positions. Divergence at first, second, and third codon positions was calculated with a custom Perl script (see “Data availability”).

From the above-mentioned bam files, haplotype calling was performed, and gVCFs were merged using GATK v4.2.6.1 (McKenna et al. 2010). Variant calling was done using the GenotypeGVCFs function in the same software, and low-quality SNVs were filtered using default parameters. These SNVs were annotated, and synonymous, nonsense, and missense mutations were identified using SnpEff v5.2a (Cingolani et al. 2012).

As in *S. frugiperda*, we identified only one protein sequence for each RefSeq gene in *S. litura* (GCF_002706865.2) and *S. exigua*

(GCA_902829305.4), and orthogroups were identified using OrthoFinder v2.5.5 (Emms and Kelly 2019). Illumina resequencing data for *S. exigua* were downloaded from NCBI SRA (accession number: ERR4094912-ERR4094927). Adaptor sequences and low-quality sequences were removed from the reads using AdapterRemoval v3.0.0 (Schubert et al. 2016). Reads were mapped against the reference genome of *S. exigua* (NCBI Genome Accession: GCA_902829305.4) using bowtie v2.5.1 (Langmead and Salzberg 2012), and potential PCR or optical duplicates were removed using picard v2.9.4 (Broad Institute). Haplotype calling was performed using GATK v4.2.0.0 (McKenna et al. 2010), followed by merging the resulting gvcf files into one. Then, variants were called using the GenotypeGVCFs function. Only SNVs were selected from the resulting vcf file, and low-quality SNVs were discarded using the default option. SNVs with the proportion of the called genotype lower than 0.8 were also discarded using vcftools v0.1.16 (Danecek et al. 2011). Synonymous and missense mutations were identified from these SNVs using SnpEff v5.2a (Cingolani et al. 2012), and pN/pS ratio was calculated from the numbers of the synonymous and missense mutations.

Regulatory sequences of iORF-CNVs

Pairs of the 300 bp upstream regions of iORF-CNVs from a single RefSeq gene were aligned using prank v170427 (Löytynoja 2014). Nucleotide *P*-distances were calculated using a custom Perl script (please see the script availability statement). The lowest distance was recorded for each iORF-CNV copy, and the histogram of these distances was generated using R v4.1.2 (<https://www.r-project.org/>).

Transcriptome data from *S. frugiperda* samples collected in Florida (Orsucci et al. 2022) were downloaded from the European Nucleotide Archive (Project ID: PRJEB25159). Adapters and low-quality reads were removed using AdapterRemoval v2.1.7 (Schubert et al. 2016). Cleaned reads were mapped to the ver7 reference genome using bowtie2 v2.4.4 (Langmead and Salzberg 2012), and read depth was calculated with samtools v1.17 (Li et al. 2009). TATA-box motifs were identified using Perl regular expressions ($\sim$ /TATA[A|T]A[A|T]/).

Synteny analysis

Conserved synteny between *S. frugiperda*, *S. litura*, and *S. exigua* was inferred using DAGchainer r02-06-2008 (Haas et al. 2004) with the protein sequences used to identify orthogroups. More specifically, these protein sequences were “blastp”ed using BLAST 2.12.0+ (Camacho et al. 2009), and the position of corresponding genes was recorded. Then, the run_DAG_chainer.pl script within the DAGchainer package was executed to identify the position of orthologous gene pairs between the two species among the three species. The BUSCO lepidoptera_odb10 proteins were “blastp”ed against the used protein sequences of *S. frugiperda*, and the position of the BUSCO genes was identified from the DAGchainer result. Then, the orthologous chromosomes among the three species were identified. The positions of orthologous BUSCO genes were visualized using the circlize v0.4.17 (Gu et al. 2014). We identified the position of intronless genes to list the deviation from the conserved synteny.

Acknowledgments

We are grateful to the Genotoul bioinformatics platform Toulouse Occitanie (Bioinfo Genotoul, <https://doi.org/10.15454/1.5572369328961167E12>). We also thank Robert L. Meagher, Jr, for his support in field sampling.

Author contributions

K.N. conceived the study, developed the experimental plans, and performed bioinformatics analyses. H.A. conducted variant annotation. F.L. found the association between contigs and chromosomes. S.G. prepared high-molecular-weight genomic DNA. E.A. and M.G. supported molecular experiments. K.D. performed gene ontology analysis. D.G.H., E.A., S.H., and S.Y. contributed to sample acquisition, including field collection, transportation, and laboratory rearing. J.B. provided the maize field for sample collection. K.N. wrote the manuscript with input from all authors.

Supplementary material

Supplementary material is available at *Molecular Biology and Evolution* online.

Funding

The study is supported by Agence Nationale De La Recherche (ORIGINS, ANR-20-CE92-0018-01) and Deutsche Forschungsgemeinschaft (HA 8327/1-1 and ORIGINS, HA 8327/2-1). This work was also supported by a grant (project no. PQ20224B013) from the Animal and Plant Quarantine Agency, Ministry of Agriculture, Food and Rural Affairs, Republic of Korea. D.G.H. was supported by the Max-Planck-Gesellschaft.

Conflicts of interest

None declared.

Data availability

Raw HiFi resequencing data and individually assembled genomes generated in this study are available at NCBI (Project Number: PRJNA1269917). Individually assembled genomes and gene annotations are also available at the Bioinformatics Platform for Agroecosystem Arthropods (https://bipaa.genouest.org/sp/spodoptera_frugiperda_pub/download/genome/corn/v7.0/). All computer programming scripts used in this study are available at GitHub (https://github.com/kiwoong-nam/sfrugi_WG).

References

Alonge M *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* 2022;23:258. <https://doi.org/10.1186/s13059-022-02823-7>.

Auton A *et al.* A global reference for human genetic variation. *Nature.* 2015;526:68–74. <https://doi.org/10.1038/nature15393>.

Balabanidou V, Grigoraki L, Vontas J. Insect cuticle: a critical determinant of insecticide resistance. *Curr Opin Insect Sci.* 2018;27:68–74. <https://doi.org/10.1016/j.cois.2018.03.001>.

Belay DK, Huckaba RM, Foster JE. Susceptibility of the fall armyworm, *Spodoptera frugiperda* (Lepidoptera: Noctuidae), at Santa Isabel, Puerto Rico, to different insecticides. *Fla Entomol.* 2012;95:476–478. <https://doi.org/10.1653/024.095.0232>.

Boen J, Wagner JP, Nanni ND. 2021 Oct 19. Inferring copy number variation from gene expression data: methods, comparisons, and applications to oncology [preprint]. *Biorxiv* 463991. <https://doi.org/10.1101/2021.10.18.463991v1>.

Broad Institute. Picard Tools. <https://broadinstitute.github.io/picard/>, accessed March 14, 2024.

Calla B *et al.* Cytochrome P450 diversification and hostplant utilization patterns in specialist and generalist moths: birth, death and adaptation. *Mol Ecol.* 2017;26:6021–6035. <https://doi.org/10.1111/mec.14348>.

Camacho C *et al.* BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10:421. <https://doi.org/10.1186/1471-2105-10-421>.

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

Charlesworth B. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet Res.* 1996;68:131–149. <https://doi.org/10.1017/S0016672300034029>.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods.* 2021;18:170–175. <https://doi.org/10.1038/s41592-020-01056-5>.

Cingolani P *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin).* 2012;6:80–92. <https://doi.org/10.4161/fly.19695>.

Conesa A, Götz S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics.* 2008;2008:619832. <https://doi.org/10.1155/2008/619832>.

Cubry P, Vigouroux Y, François O. The empirical distribution of singletons for geographic samples of DNA sequences. *Front Genet.* 2017;8:139. <https://doi.org/10.3389/fgene.2017.00139>.

Danecek P *et al.* The variant call format and VCFtools. *Bioinformatics.* 2011;27:2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>.

Day R *et al.* Fall armyworm: impacts and implications for Africa. *Outlooks Pest Manag.* 2017;28:196–201. https://doi.org/10.1564/v28_oct_02.

Donczew R, Hahn S. Mechanistic differences in transcription initiation at TATA-less and TATA-containing promoters. *Mol Cell Biol.* 2017;38:e00448-17. <https://doi.org/10.1128/MCB.00448-17>.

Dong S *et al.* Identification of chorion genes and RNA interference-mediated functional characterization of *chorion-1* in *Plutella xylostella*. *J Integr Agric.* 2022;21:3278–3292. <https://doi.org/10.1016/j.jia.2022.08.079>.

Dort H, van der Bijl W, Wahlberg N, Nylin S, Wheat CW. Genome-wide gene birth-death dynamics are associated

- with diet breadth variation in lepidoptera. *Genome Biol Evol.* 2024;16:evae095. <https://doi.org/10.1093/gbe/evae095>.
- Doucet AJ, Wilusz JE, Miyoshi T, Liu Y, Moran JV. A 3' Poly(A) tract is required for LINE-1 retrotransposition. *Mol Cell.* 2015;60:728–741. <https://doi.org/10.1016/j.molcel.2015.10.012>.
- Durand K, An H, Nam K. Invasive fall armyworms are corn strain. *Sci Rep.* 2024;14:5696. <https://doi.org/10.1038/s41598-024-56301-0>.
- Durand K, Nam K. Is the extreme within-population genome size variation real in *Spodoptera frugiperda*? *PLoS One.* 2025;20:e0332711. <https://doi.org/10.1371/journal.pone.0332711>.
- Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science.* 2008;320:1629–1631. <https://doi.org/10.1126/science.1158078>.
- Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019;20:238. <https://doi.org/10.1186/s13059-019-1832-y>.
- Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet.* 2000;24:363–367. <https://doi.org/10.1038/74184>.
- Eyre-Walker A, Woolfit M, Phelps T. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics.* 2006;173:891–900. <https://doi.org/10.1534/genetics.106.057570>.
- Fang B, Edwards SV. Fitness consequences of structural variation inferred from a House Finch pangenome. *Proc Natl Acad Sci U S A.* 2024;121:e2409943121. <https://doi.org/10.1073/pnas.2409943121>.
- Fiteni E *et al.* Host-plant adaptation as a driver of incipient speciation in the fall armyworm (*Spodoptera frugiperda*). *BMC Ecol Evol.* 2022;22:133. <https://doi.org/10.1186/s12862-022-02090-x>.
- Flynn JM *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A.* 2020;117:9451–9457. <https://doi.org/10.1073/pnas.1921046117>.
- Gao X, Li Y, Adetula AA, Wu Y, Chen H. Analysis of new retrogenes provides insight into dog adaptive evolution. *Ecol Evol.* 2019;9:11185–11197. <https://doi.org/10.1002/ece3.5620>.
- Gimenez S *et al.* Adaptation by copy number variation increases insecticide resistance in the fall armyworm. *Commun Biol.* 2020;3:664. <https://doi.org/10.1038/s42003-020-01382-6>.
- Girirajan S, Campbell CD, Eichler EE. Human copy number variation and complex genetic disease. *Annu Rev Genet.* 2011;45:203–226. <https://doi.org/10.1146/annurev-genet-102209-163544>.
- Goergen G, Kumar PL, Sankung SB, Togola A, Tamò M. First report of outbreaks of the Fall Armyworm *Spodoptera frugiperda* (J E Smith) (Lepidoptera, Noctuidae), a new alien invasive pest in west and Central Africa. *PLoS One.* 2016;11:e0165632. <https://doi.org/10.1371/journal.pone.0165632>.
- Goodier JL, Ostertag EM, Kazazian HH. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum Mol Genet.* 2000;9:653–657. <https://doi.org/10.1093/hmg/9.4.653>.
- Gouin A *et al.* Two genomes of highly polyphagous lepidopteran pests (*Spodoptera frugiperda*, Noctuidae) with different host-plant ranges. *Sci Rep.* 2017;7:11816. <https://doi.org/10.1038/s41598-017-10461-4>.
- GraphicsMagick Group. GraphicsMagick Image Processing System. <http://www.graphicsmagick.org/>, accessed April 10, 2026.
- Grzybowska EA. Human intronless genes: functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. *Biochem Biophys Res Commun.* 2012;424:1–6. <https://doi.org/10.1016/j.bbrc.2012.06.092>.
- Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. *Bioinforma. Oxf. Engl.* 2014;30:2811–2812. <https://doi.org/10.1093/bioinformatics/btu393>.
- Haas BJ, Delcher AL, Wortman JR, Salzberg SL. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics.* 2004;20:3643–3646. <https://doi.org/10.1093/bioinformatics/bth397>.
- Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet.* 2009;10:551–564. <https://doi.org/10.1038/nrg2593>.
- Hoile AE, Holland PWH, Mulhair PO. Gene novelty and gene family expansion in the early evolution of Lepidoptera. *BMC Genomics.* 2025;26:161. <https://doi.org/10.1186/s12864-025-11338-x>.
- Hughes AL. Adaptive evolution after gene duplication. *Trends Genet.* 2002;18:433–434. [https://doi.org/10.1016/S0168-9525\(02\)02755-5](https://doi.org/10.1016/S0168-9525(02)02755-5).
- Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 2010;11:97–108. <https://doi.org/10.1038/nrg2689>.
- Jayakodi M *et al.* Structural variation in the pangenome of wild and domesticated barley. *Nature.* 2024;636:654–662. <https://doi.org/10.1038/s41586-024-08187-1>.
- Jiang H. 2026. moold/paf2dotplot. <https://github.com/moold/paf2dotplot>, accessed April 10, 2026.
- Keightley PD *et al.* Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Mol Biol Evol.* 2015;32:239–243. <https://doi.org/10.1093/molbev/msu302>.
- Kim S *et al.* Transposable element-mediated structural variation analysis in dog breeds using whole-genome sequencing. *Mamm Genome.* 2019;30:289–300. <https://doi.org/10.1007/s00335-019-09812-5>.
- Lai H-Y *et al.* iProEP: a computational predictor for predicting promoter. *Mol Ther Nucleic Acids.* 2019;17:337–346. <https://doi.org/10.1016/j.omtn.2019.05.028>.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9:357–359. <https://doi.org/10.1038/nmeth.1923>.
- Lefouili M, Nam K. The evaluation of Bcftools mpileup and GATK HaplotypeCaller for variant calling in non-human species. *Sci Rep.* 2022;12:11331. <https://doi.org/10.1038/s41598-022-15563-2>.
- Li H *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics.* 2009;25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.
- Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018;34:3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
- Li H. Protein-to-genome alignment with miniprot. *Bioinformatics.* 2023;39:btad014. <https://doi.org/10.1093/bioinformatics/btad014>.
- Li Z *et al.* The pig pangenome provides insights into the roles of coding structural variations in genetic diversity and

- adaptation. *Genome Res.* 2023;33:1833–1847. <https://doi.org/10.1101/gr.277638.122>.
- Löytynoja A. Phylogeny-aware alignment with PRANK. *Methods Mol. Biol. Clifton NJ.* 2014;1079:155–170. https://doi.org/10.1007/978-1-62703-646-7_10.
- Lye ZN, Purugganan MD. Copy number variation in domestication. *Trends Plant Sci.* 2019;24:352–365. <https://doi.org/10.1016/j.tplants.2019.01.003>.
- Macias-Muñoz A, McCulloch KJ, Briscoe AD. Copy number variation and expression analysis reveals a nonorthologous Pinta gene family member involved in butterfly vision. *Genome Biol Evol.* 2017;9:3398–3412. <https://doi.org/10.1093/gbe/evx230>.
- Maizels N. Immunoglobulin gene diversification. *Annu Rev Genet.* 2005;39:23–46. <https://doi.org/10.1146/annurev.genet.39.073003.110544>.
- McKenna A et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–1303. <https://doi.org/10.1101/gr.107524.110>.
- Montezano DG et al. Host plants of *Spodoptera frugiperda* (Lepidoptera: Noctuidae) in the Americas. *Afr Entomol.* 2018;26:286–300. <https://doi.org/10.4001/003.026.0286>.
- Moran JV, DeBerardinis RJ, Kazazian HH. Exon shuffling by L1 Retrotransposition. *Science.* 1999;283:1530–1534. <https://doi.org/10.1126/science.283.5407.1530>.
- Mulhair PO, Crowley L, Boyes DH, Lewis OT, Holland PWH. Opsin gene duplication in Lepidoptera: retrotransposition, sex linkage, and gene expression. *Mol Biol Evol.* 2023;40:msad241. <https://doi.org/10.1093/molbev/msad241>.
- Nguyen AK, Blacksmith MS, Kidd JM. Duplications and retrogenes are numerous and widespread in modern canine genomic assemblies. *Genome Biol Evol.* 2024;16:evae142. <https://doi.org/10.1093/gbe/evae142>.
- Ohno S. *Evolution by gene and genome duplication*. Springer; 1970.
- Orsucci M et al. Transcriptional differences between the two host strains of *Spodoptera frugiperda* (Lepidoptera: Noctuidae). *Peer Community J.* 2022;2:e1. <https://doi.org/10.24072/pjournal.77>.
- Overton K et al. Global crop impacts, yield losses and action thresholds for fall armyworm (*Spodoptera frugiperda*): a review. *Crop Prot.* 2021;145:105641. <https://doi.org/10.1016/j.cropro.2021.105641>.
- Pashley DP. Host-associated genetic differentiation in fall armyworm (Lepidoptera: Noctuidae): a sibling species complex? *Ann Entomol Soc Am.* 1986;79:898–904. <https://doi.org/10.1093/aesa/79.6.898>.
- Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics.* 2018;34:867–868. <https://doi.org/10.1093/bioinformatics/btx699>.
- Pertea G, Pertea M. GFF utilities: GffRead and GffCompare. *F1000Res.* 2020;9:ISCB Comm J-304. <https://doi.org/10.12688/f1000research.23297.2>.
- Pös O et al. DNA copy number variation: main characteristics, evolutionary significance, and pathological aspects. *Biomed J.* 2021;44:548–559. <https://doi.org/10.1016/j.bj.2021.02.003>.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–842. <https://doi.org/10.1093/bioinformatics/btq033>.
- Redon R et al. Global variation in copy number in the human genome. *Nature.* 2006;444:444–454. <https://doi.org/10.1038/nature05329>.
- Richardson SR, Salvador-Palomeque C, Faulkner GJ. Diversity through duplication: whole-genome sequencing reveals novel gene retrocopies in the human population. *BioEssays.* 2014;36:475–481. <https://doi.org/10.1002/bies.201300181>.
- Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods.* 2020;17:155–158. <https://doi.org/10.1038/s41592-019-0669-3>.
- Schloissnig S et al. 2025. Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature.* 2025;644:442–452. <https://doi.org/10.1038/s41586-025-09290-7>.
- Schrider DR et al. Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet.* 2013;9:e1003242. <https://doi.org/10.1371/journal.pgen.1003242>.
- Schrider DR, Hahn MW. Gene copy-number polymorphism in nature. *Proc R Soc B Biol Sci.* 2010;277:3213–3221. <https://doi.org/10.1098/rspb.2010.1180>.
- Schrider DR, Stevens K, Cardeño CM, Langley CH, Hahn MW. Genome-wide analysis of retrogene polymorphisms in *Drosophila melanogaster*. *Genome Res.* 2011;21:2087–2095. <https://doi.org/10.1101/gr.116434.110>.
- Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes.* 2016;9:88. <https://doi.org/10.1186/s13104-016-1900-2>.
- Shao X et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. *BMC Med Genet.* 2019;20:175. <https://doi.org/10.1186/s12881-019-0909-5>.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 2015;31:3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>.
- Sirén J et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science.* 2021;374:abg8871. <https://doi.org/10.1126/science.abg8871>.
- Smit AFA, Hubley R, Green P. RepeatMasker. *RepeatMasker Open-40*, 2013. <http://www.repeatmasker.org/>, accessed December 17, 2024.
- Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. *Genet Res.* 1974;23:23–35. <https://doi.org/10.1017/S0016672300014634>.
- Spealman P, de Santana C, De T, Gresham D. Multilevel gene expression changes in lineages containing adaptive copy number variants. *Mol Biol Evol.* 2025;42:msaf005. <https://doi.org/10.1093/molbev/msaf005>.
- Sudmant PH et al. Diversity of human copy number variation and multicopy genes. *Science.* 2010;330:641–646. <https://doi.org/10.1126/science.1197005>.
- Vaser R, Šikić M. Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci.* 2021;1:332–336. <https://doi.org/10.1038/s43588-021-00073-4>.
- Vaser R, Sović I, Nagarajan N, Sikic M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 2017;27:737–746. <https://doi.org/10.1101/gr.214270.116>.
- Wang B et al. Cuticle thickening mediates insecticide penetration resistance in *Spodoptera litura*. *J Adv Res.* 2025;78:163–178. <https://doi.org/10.1016/j.jare.2025.02.027>.

- Wright CJ, Stevens L, Mackintosh A, Lawniczak M, Blaxter M. Comparative genomics reveals the dynamics of chromosome evolution in Lepidoptera. *Nat Ecol Evol.* 2024;8:777–790. <https://doi.org/10.1038/s41559-024-02329-4>.
- Yainna S *et al.* The evolutionary process of invasion in the fall armyworm (*Spodoptera frugiperda*). *Sci Rep.* 2022;12:21063. <https://doi.org/10.1038/s41598-022-25529-z>.
- Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. *Nat Rev Genet.* 2015;16:172–183. <https://doi.org/10.1038/nrg3871>.
- Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet.* 2009;10:451–481. <https://doi.org/10.1146/annurev.genom.9.081307.164217>.
- Zhang W, Tautz D. Tracing the origin and evolutionary fate of recent gene retrocopies in natural populations of the house mouse. *Mol Biol Evol.* 2022;39:msab360. <https://doi.org/10.1093/molbev/msab360>.
- Zhang W, Xie C, Ullrich K, Zhang YE, Tautz D. The mutational load in natural populations is significantly affected by high primary rates of retroposition. *Proc Natl Acad Sci U S A.* 2021;118:e2013043118. <https://doi.org/10.1073/pnas.2013043118>.
- Zhengcao L *et al.* The pig pangenome provides insights into the roles of coding structural variations in genetic diversity and adaptation. *Genome Res.* 2023;33:1833–1847. <https://doi.org/10.1101/gr.277638.122>.
- Zhu J-Y, Li L, Xiao K-R, He S-Q, Gui F-R. Genomic and transcriptomic analysis reveals cuticular protein genes responding to different insecticides in Fall Armyworm *Spodoptera frugiperda*. *Insects.* 2021;12:997. <https://doi.org/10.3390/insects12110997>.