**Title:** Cryptic diversity and impacts of domestication in the Black Soldier Fly (*Hermetia illucens*) genome

**Short Title:** Genetic diversity in the Black Soldier Fly

Tomas N. Generalovic[1]*, Christoph Sandrock[2], Benjamin J. Roberts[3], Joana I. Meier[1,4], Martin Hauser[5], Ian A. Warren[1], Miha Pipan[6], Richard Durbin[7] & Chris D. Jiggins[1].

[1]Department of Zoology, University of Cambridge, Cambridge, UK; [2]Department of Livestock Sciences, Research Institute of Organic Agriculture (FiBL), Frick, Switzerland; [3]Georgina Mace Centre for the Living Planet, Faculty of Natural Sciences, Imperial College London, London, UK; [4]Tree of Life Programme, Wellcome Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK; [5]California Department of Food and Agriculture, Plant Pest Diagnostics Branch, Sacramento, CA, USA; [6]Better Origin, Entomics Biosystems Limited, Cambridge, UK; [7]Department of Genetics, University of Cambridge, Cambridge, UK

*Corresponding Author: Tomas N. Generalovic. Email: tng23@cam.ac.uk

## Abstract

The Black Soldier Fly (*Hermetia illucens*) is the main species in the developing global industry of insects as food and feed, but little is known about its natural diversity or the genetic basis of its domestication. We obtained whole genome sequences for 54 individuals from both wild and captive populations. We identified two genetic lineages at least 3 million years divergent, revealing cryptic diversity within the species. Our study indicates that the most common populations used for commercial and academic applications are primarily derived from just one of these lineages, likely originating from a wild North American progenitor. We find that captive populations show strong reductions in genetic diversity, consistent with genome-wide effects of population bottlenecks and drift associated with rearing in captivity. Some limited evidence of gene flow between divergent lineages was observed, as well as evidence of hybridisation from domesticated populations into the wild. Our study suggests that natural genetic diversity could provide important variation for industrial purposes in this novel agricultural species.

**Author Summary**

Animals and plants have been domesticated by humans over thousands of years, shaping their genomes and yielding beneficial or improved traits for agriculture. Insect farming has recently reached industrial scales in order to recycle organic wastes into feed and food grade material. The Black Soldier Fly is the flagship insect species in this novel agriculture sector. However, the understanding of its evolutionary history and the genetic impact of intense farming has not been well studied to date. We examined whole genomes from populations sampled from across the worldwide distribution of Black Soldier Fly to investigate the impact of domestication and reveal the evolutionary history of the species. We found that farmed populations show distinct genomic regions showing reduced variation that could represent signatures of selection or demographic effects. We also identify two distinct lineages of Black Soldier Fly that have been diverging for millions of years, comparable to divergence between different species in other insect groups, yet show evidence of genetic exchange, indicating a lack of reproductive incompatibility in the populations tested. This work is the first to describe the genetic impact of domestication in the Black Soldier Fly and provide evolutionary insights using whole genomes. These results imply that considerable wild diversity could be harnessed for improving agricultural efficiency.

## Introduction

Domestication of plants and animals has enabled successful human population growth and globalisation [1]. Domestication is a coevolutionary process that emerges from a specialised mutualism where a species fitness is managed through manipulation of the environment, food-source, behaviour, and mate choice [2,3]. Evolutionary changes in physiological, morphological, and behavioural traits over time have led to specialised crops and animals with a host of phenotypic diversity, exemplified by *Brassica oleracea* cultivars and dog breeds [4,5]. This process of domestication also provides insights into evolution on short time scales [6,7]. Whilst domestication has been extensively studied in plants, mammals, birds and fish [4,5,8,9], the silkworm, *Bombyx mori*, is currently the only undisputed domesticated insect [10]. Here we focus on another candidate domesticated insect, the Black Soldier Fly, *Hermetia illucens* (L., 1758; Diptera: Stratiomyidae), in which both captive and wild populations can be found around the globe.

The Black Soldier Fly has only become an important and globally mass farmed species within the past two decades [11–13]. In response to rising food demands driven by an ever-increasing human population, novel and more sustainable agricultural practices are being developed, including the insects for food and feed industry [13,14]. The Black Soldier Fly can convert a diverse range of organic wastes into protein and fat rich biomass [11,15,16]. It has a cosmopolitan distribution with an origin in the Americas [17], most likely South America [18]. The global distribution of the Black Soldier Fly outside of the Americas is thought to be a result of human mediated dispersal and by 1960 the species had established most of its present-day distribution [17]. Microsatellite and mtDNA studies have shown that the species harbours high levels of genetic diversity [18–20], which suggests considerable potential for artificial selection and the development of improved strains [21,22]. However, commercial activity is currently dominated by a few closely related populations all derived from one strain founded in Georgia, USA in the 1990's by Craig Sheppard [12,18].

Captive populations such as these have likely undergone or are currently experiencing a multitude of domestication bottlenecks and genetic adaptions associated with artificial insect farming conditions.

Whilst previous studies using microsatellites and mtDNA have provided initial insights into genetic diversity in Black Soldier Fly [18–20], these markers only sample a small proportion of the genome providing an incomplete view of genomic diversity [23]. In contrast, whole-genome sequencing captures variation genome-wide providing a comprehensive view of population diversity and structure [24]. Whole-genome sequencing is therefore able to provide more accurate estimates of genetic diversity, differentiation and patterns of linkage disequilibrium. Importantly, this enables genome wide patterns of drift and selection, signals that are typically undetectable with microsatellite data, to be detected. These features make whole-genome sequencing well-suited for revealing the evolutionary processes during domestication and within-species divergence.

Here we explore Black Soldier Fly diversity on a whole-genome scale, using a previously published high quality reference genome [25]. Using this first genome-wide dataset of the Black Soldier Fly, sequenced to high coverage, we explore the genomic landscape of captive and wild populations from key geographic locations. We address the following questions: (1) What are the phylogenetic relationships among Black Soldier Fly lineages? (2) When comparing genomes of captive populations to wild relatives can we reveal common patterns associated with domestication? (3) What are the genetic consequences of introgression in the wild and within insect mass-rearing commercial facilities?

## Results

### Genome sequencing and variant calling

We sequenced 54 Black Soldier Fly genomes to an average 22x coverage alongside an additional outgroup species, *Ptecticus aurifer* (supplementary table 1). To capture the extent

of genetic variation within the species we sampled Black Soldier Fly individuals from genetically distinct (captive and wild) populations, previously identified from microsatellite data [18]. The remaining dataset consisted of representative samples of active industrial and academic captive lines. In addition, we sampled wild sourced populations as part of a taxonomic voucher set including a single outgroup species. We hereafter term academic/industry-derived samples as 'captive' (*n*=35/54) whereas samples directly sampled from in the field were termed 'wild' (*n*=19/54) (supplementary table 1). Captive individuals belonging to South America, Africa and Australia were recorded as wild caught and recently established as a captive colony. After variant calling and stringent quality filtering, we identified 33,940,069 bi-allelic Single Nucleotide Polymorphisms (SNPs).

**Divergent lineages and cryptic species complex**

Phylogenetic analysis on the genome-wide SNP data showed two major genetic lineages. The first lineage contains wild sampled individuals from North America and many captive populations, hereafter named lineage-*a* (Figure 1A). The captive populations found in lineage-*a* are made polyphyletic by the North American wild population. European and North American captive populations form sister clades in line with a documented split from an original founding population, named hereafter as the 'Sheppard' strain. The 'EVE' genome reference line also appears to be a derivative of the Sheppard strain. The second lineage, lineage-*b,* is formed from a diverse range of wild and captive populations. In lineage-*b*, wild and captive clades broadly cluster according to geographic location, indicating these captive populations are established from locally occurring populations, as documented. Individuals from Asia (wild) and Australia (captive) also form an 'Australasian' clade. Phylogenetic methods using coalescence-based methods (ASTRAL) with input trees calculated using either Neighbour-Joining (NJ) or Maximum-Likelihood (ML) algorithms also support a split between major lineages (supplementary figure 1A & 1B). However, the NJ approach

suggests lineage-*b* samples are paraphyletic with respect to lineage-*a* individuals (supplementary figure 1A).

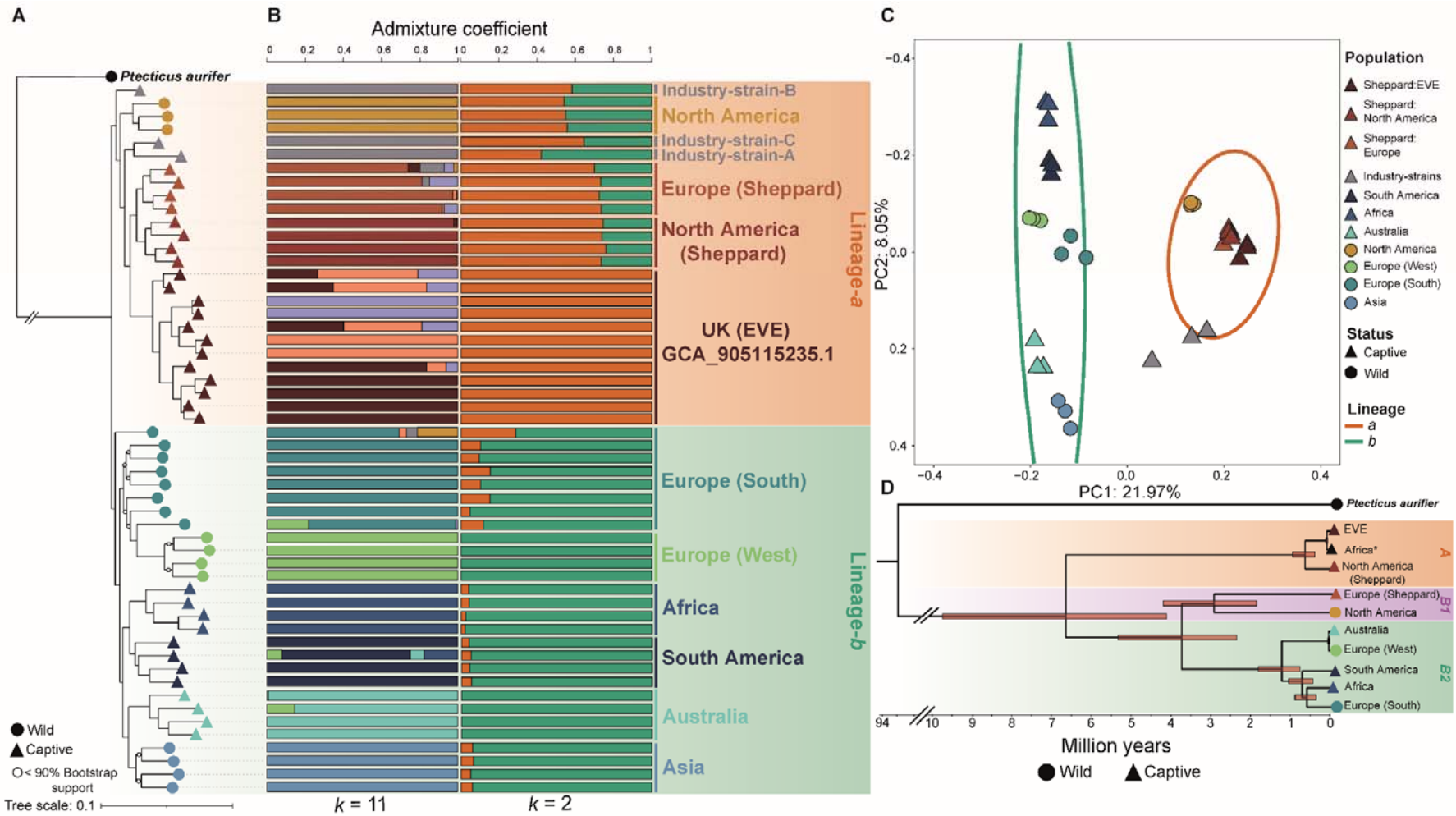


**Figure 1**. (**A**) Maximum Likelihood phylogenetic reconstruction using whole Black Soldier Fly (n=54) and *Ptecticus aurifer* (n=1) genomes. (**B**) ADMIXTURE analysis of the Black Soldier Fly population structure at *k*=11 and *k*=2 (lowest cross-validation error support) indicating pronounced population structure. (**C**) Principal Component Analysis (PCA) using a subsampled dataset of whole genomes (n=3 per population) supporting two discrete genetic lineages. A 95% confidence interval (CI) ellipse is drawn over major lineages based on phylogenetic (**A**) and clustering (**B**; *k*=2) assignment of individuals. (**D**) BEAST time divergence model representation of the Black Soldier Fly mitochondrial phylogeny reconstructed using a single whole mitochondrial (mt) genome from each major clade identified in the whole-genome phylogeny (**A**). Lineages recovered in the mtDNA tree are labelled including the unresolved mtDNA lineage-*B1*. *Single mt genome obtained from [20].

Evidence for substantial divergence between the two nuclear lineages was also supported by ADMIXTURE (Figure 1B) and Principal Component Analysis (PCA; Figure 1C). Optimal clustering in ADMIXTURE analysis was identified at $k$ = 2 (Figure 1B; supplementary table 2), and this was consistent when subsampling an equal number of individuals per population (supplementary figure 2). A similar Cross Validation (CV) rate was identified at $k$=3 (0.887) to $k$=2 (0.885) indicating a third clustering group primarily consisting of EVE and captive North American individuals (supplementary figure 3). At higher $k$-values, $k$ = 11 (Figure 1B), wild European populations show distinct clusters indicating a diverse within continent population structure consistent with independent colonisation events and likely subsequent patterns of isolation-by-distance. A similar pattern is observed across a larger geographic range in Asian and Australian samples in agreement with the successive colonisation of Australasia [18,20]. PCA analysis also showed two main clusters with many captive populations clustering with wild North American samples while the remaining wild and captive samples found from a broad geographic range form a second major cluster (Figure 1C). Several individuals belonging to various industry populations (Ind-A, -B & -C) are found in the PCA as intermediates between lineages, suggesting admixed origins for these populations. When grouped by discrete populations identified in clustering analysis we found moderate genome-wide levels of $F_{ST}$ (0.178) and $d_{XY}$ (0.081) between nuclear lineages. This suggests a long history of evolutionary separation with limited gene flow between lineages. Within lineages, lineage-*b* had higher genome-wide values of both $F_{ST}$ and $d_{XY}$ as compared to lineage-*a* (supplementary figure 4).

**Ancient divergence between Black Soldier Fly lineages**

We produced *de novo* mitochondrial assemblies for the 54 Black Soldier Fly individuals and one *P. aurifer* individual (supplementary table 3; supplementary figure 5). The *de novo* Black Soldier Fly and *P. aurifer* assemblies are comparable to published reference mitochondrial genomes (supplementary table 3; supplementary figure 6) [20,26,27]. In contrast to the two

lineages (*a* & *b*) identified from nuclear sequences, three major lineages were recovered in the mtDNA data (*A*, *B1* & *B2*) (Figure 1D & supplementary figure 6). Nuclear lineage-*a* samples fell into mitochondria clades *A* and *B1*, in which lineage-*A* diverged an estimated 4.1 to 9.7 million years ago (my; 95% Confidence Intervals (CI); mean 6.8 my) (supplementary table 4). Lineage-*B1* of the mtDNA tree contains nuclear lineage-*a* wild North American and captive European-Sheppard populations (Figure 1D). mtDNA lineage-*B2* contains all samples from nuclear lineage-*b* and diverged from lineage-*B1* approximately 2.4 – 5.3 my ago (95% CI; mean 3.8 my) (supplementary table 4). Confidence in the models dating outputs is high, in the initial calibrating dating run, comprising a wider dipteran phylogeny, all calibrated nodes fit well within the assigned constraints whilst non-calibrated nodes show agreement with existing literature (supplementary table 4), suggesting that the dating model has performed well. Here, we dated the Stratiomyidae - Tabanidae split at 182.0 my ago and the Stratiomyidae root at 81.1 my whilst these splits have been previously dated at around ~190 my and ~75 my ago respectively (see supplementary table 4 and references within), well within the 95% confidence intervals. We also generated Neighbour-Joining (NJ) trees of Cytochrome Oxidase I (COI) markers, such that additional previously published samples could be incorporated, including some from wild Central American populations (supplementary figure 7). This confirms that Sheppard-derived populations of lineage-*a* are most closely related to wild samples from North America.

To corroborate the divergence times between lineages identified in the mtDNA data we performed additional dating analysis using whole genome evidence. Using the known split, between *P. aurifer* and *H. illucens,* we constrained the phylogeny to predict the divergence between nuclear lineages-*a* and -*b*. Our whole-genome analysis predicts that the major split between lineages was estimated between 3.9 – 13.8 my ago (95% CI; mean 8.9 my) (supplementary figure 8). Whilst the mean whole-genome divergence between lineages is higher, in both mtDNA analysis provided above and in published work [20], the majority of

the range (3.9 -13.8 my) overlaps considerably with the mtDNA confidence intervals of 4.1 - 9.7 my and further suggests considerable diversity in the Black Soldier Fly.

**Genetic signals of domestication**

Genomic signals of domestication are typically characterised by genome-wide reductions in genetic diversity and elevated linkage disequilibrium (LD) associated with a domestication bottleneck and small effective population sizes (*Ne*). On average, genome-wide heterozygosity of captive Sheppard (0.031) strains shows a 39% reduction compared to wild populations (0.051), supporting the domestication hypothesis (supplementary table 5; supplementary figure 9). Additional captive populations from South America exhibit the highest levels of diversity (0.087), as expected if this represents the ancestral range of the species despite captivity status, followed by African populations. In general, extensively reared captive populations showed lower diversity with the exception of the European population of the Sheppard strain which showed levels of diversity similar to that of wild populations. Within-population estimates of LD revealed elevated levels in captive populations consistent with a genetic bottleneck during domestication (supplementary figure 10). This analysis highlighted genome-wide patterns of domestication in captive populations.

Due to novel selection pressures and somewhat recent bottlenecking during domestication, genomes of captive populations would be expected to contain higher frequencies of homozygous genomic regions, due to a loss of rare alleles, compared to the wild. We generated evidence for genomic changes associated with domestication and demographic events by identifying regions of reduced diversity by calculating Runs of Homozygosity (RoH) (supplementary table 6 & 7). In total, captive populations (*n* = 5,245) contained a higher number of RoH than wild (*n* = 352) based populations. However, to quantify unique RoH we reduced overlapping regions into single counts that were present in all individuals within captivity status groups. We identified a total of 907 captive and 53 wild unique RoH with 245 shared RoH regions (supplementary figure 11). Although only 53 RoH

regions were unique to wild populations, these accounted for just 6.5 Mbp, whilst 80.9 Mbp (83%) of the wild RoH regions were shared with captive populations whereas captive populations span 623.0 Mbp of the genome. Captive RoH were significantly more numerous (Wilcoxon rank-sum test, $W = 445$, $p < 0.001$) and on average longer (Wilcoxon rank-sum test, $W = 444$, $p < 0.001$) at 0.69 Mb, indicating recent bottlenecks, compared to both shared (0.33 Mb) and wild (0.12 Mb) populations. We also performed a permutation test based on 10,000 randomisations across the genome to assess if the total unique captive RoH regions had more or less coverage than expected by random across the genome. Interestingly by assessing the unique captive RoH regions we found significantly lower than expected ($p < 0.001$, $Z = -45.4$). Whilst numerous and long, this result suggests that captive RoH's appear tightly clustered and less dispersed across the genome indicating hotspots impacted by demographic or selection processes (supplementary table 8). However, many of these patterns are somewhat driven by extensive RoH within the EVE population. We identified six regions in which RoH clustered consistently in at least 45% of the captive individual samples (Figure 2).

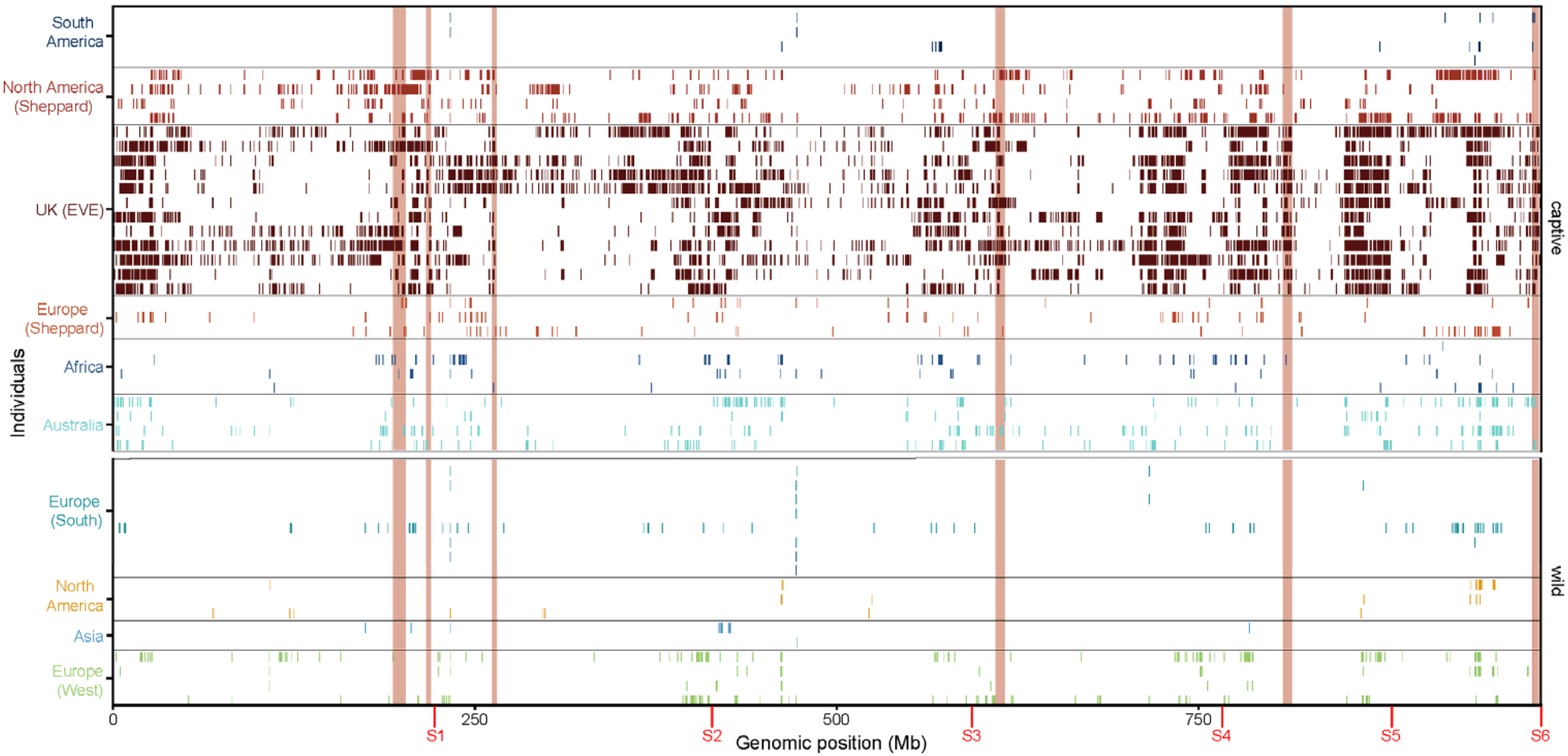


**Figure 2.** Runs of Homozygosity calculated across the genomes of wild and captive Black Soldier Fly populations. Each row in the y-axis represents an individual and colour represents the population. Chromosome boundaries are provided in red on the x-axis along with genome position (Mb). Regions clustered with RoH are highlighted in coloured blocks highlighting regions in which RoH cluster in over 45% of the captive individuals on chromosomes S1 (*n* =2), S2 (*n* =1), S4 (*n* =1), S5 (*n* =1) and S6 (*n*=1).

We also used genome-wide scans of genetic diversity ($\pi$) and differentiation ($F_{ST}$) to further investigate the genomes for differences between captive Sheppard populations and their closest sampled wild relative, North America. Strong patterns of genome-wide nucleotide diversity between captive and wild populations were relatively absent with North America (wild) and South America (wild) both showing higher levels of diversity whereas captive populations were reduced (supplementary figure 12). Elevated $F_{ST}$ was observed genome-wide for all three populations, however, strong peaks were observed converging on chromosome five for all three populations (Figure 3). Notably, there was strong $F_{ST}$ peaks overlapping in the Sheppard North American and European population on chromosomes one, four and five. The EVE population showed broad and elevated $F_{ST}$ peaks, for example, the large RoH locus on chromosome five previously identified in the EVE population had broadly elevated $F_{ST}$ but was not identified in other Sheppard populations suggesting this locus is likely demographic in nature and specific to the EVE population. Combined, these data suggest that domestication has left a strong signature on the genomes of captive populations, either through drift or selection.

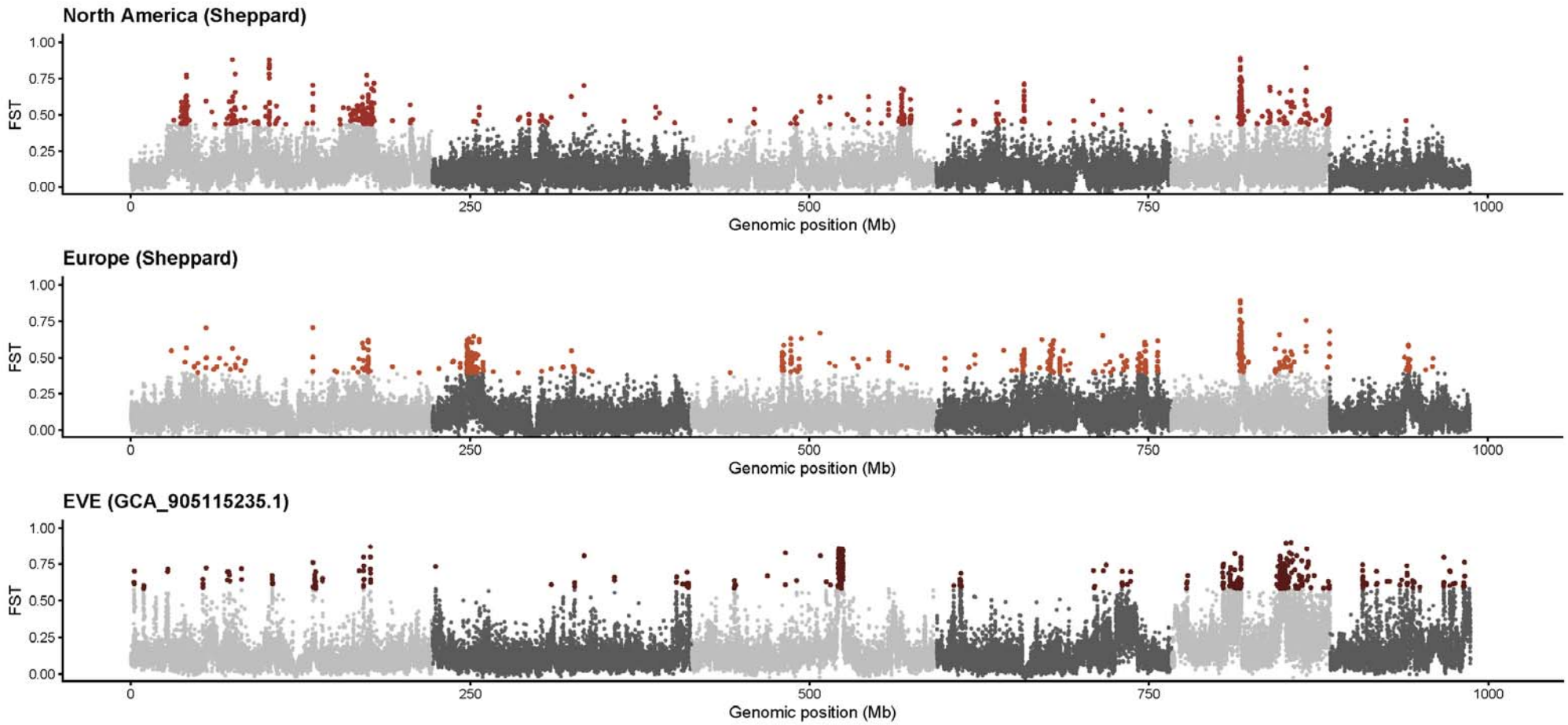


**Figure 3.** Genetic differentiation ($F_{ST}$) between various domesticated Black Soldier Fly populations. Sheppard strain from North America, Europe, and the EVE populations are compared to the closest wild sampled progenitor population from North America. Outlier regions identified within the top 99$^{th}$ percentile of 20 kb windows containing SNPs are indicated using population specific-colored points.

## Genomic consequences of introgression

Evidence for introgression between the major lineages of Black Soldier Fly comes from a comparison between nuclear (n=54) and mitochondria (n=112) whole-genome phylogenies, which showed mito-nuclear discordance (Figure 4). We found 12 samples with mito-nuclear discordance at the broad level of the major lineages and clades discussed above; all but two of these were localised to mtDNA clade-*B1*. However, overall tree discordance was low affecting only 22% of sampled individuals.

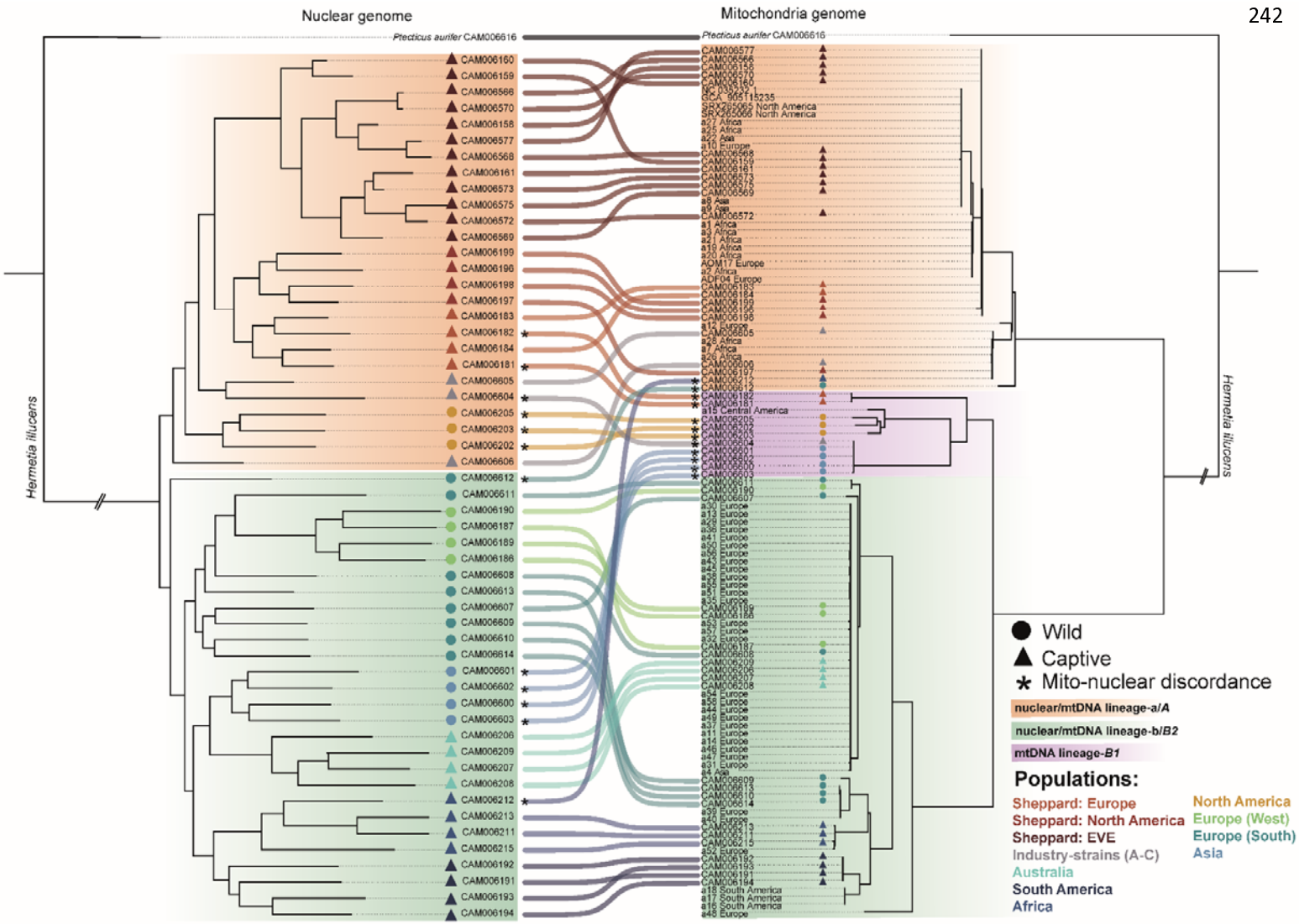


**Figure 4**. Co-phylogenetic plot of the Black Soldier Fly nuclear (left) and mitochondrial DNA (right) linked using population specific-coloured ribbons and individual CAM ID's. Evidence of mito-nuclear discordance is described by an asterisk in individuals that cross the major identified lineages (*A*, *B1* & *B2*). Whole mitochondria genomes from this study (**I**) and various (**II**) public repositories are used to expand the mtDNA tree.

We next examined genome-wide patterns of introgression within and between divergent lineages. We generated *f*-branch ($f_b$) statistics using the whole-genome phylogeny of the Black Soldier Fly to estimate genome-wide allele sharing patterns between major lineages using *P. aurifer* as the distant outgroup. Generally, there was little evidence for gene flow between the major lineages, with a few notable exceptions (Figure 5A).

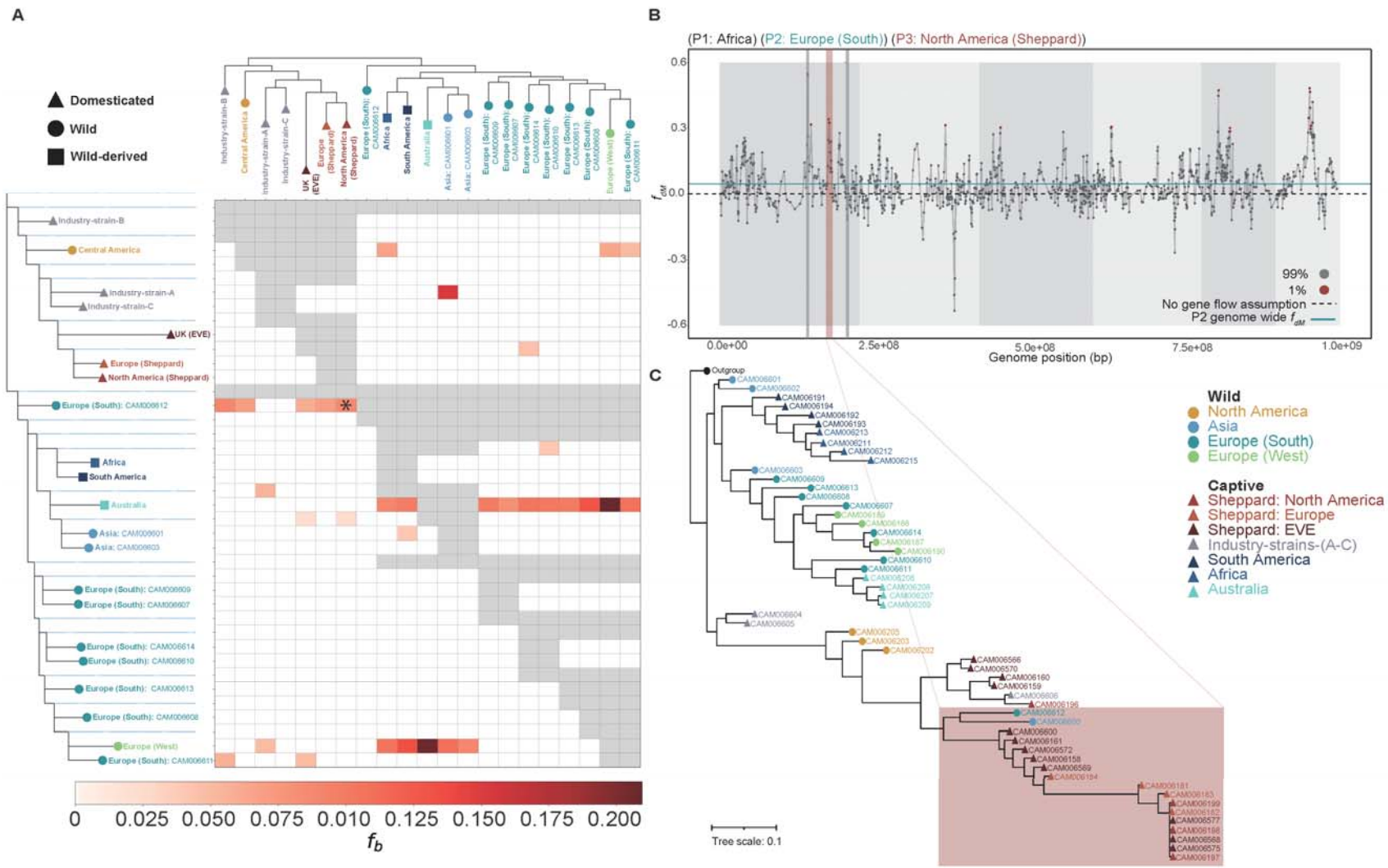


**Figure 5.** (**A**) Admixture analysis using the *f*-branch ($f_b$) statistic across sampled Black Soldier Fly individuals indicating excess allele sharing as indicated by utilised species trees (left and top). Phylogenetic tree on the y-axis represents populations P1 & P2 and is expanded to indicate internal branches enabling comparisons to the P3 population along the x-axis. Admixture is detected between P2 & P3 with P1 fixed as the closest sister group. Grid cell colour and intensity indicates levels of significant allele sharing. Grey shading indicates where the tests cannot be performed due to tree topology. Significant values only are plotted. *Indicates zoomed example for panels B & C. (**B**) Genome-wide $f_{DM}$ plotted highlighting 1% outlier values of allele sharing between P3 and P2. Alternating light and dark grey shading indicate autosomes. (**C**) Local gene tree of the introgressed region (2.1 Mb; grey) from the North American Sheppard Black Soldier Fly population (P3) into the Southern Europe wild (P2) individual showing clustering with captive populations (highlighted block).

The strongest evidence of introgression from lineage-*a* into -*b* ($f_b$ = 0.06 - 0.09) occurred from captive lineage-*a* into a wild southern European (CAM006612) representative (Figure 5A). This same individual was previously identified as showing mito-nuclear discordance (Figure 4). This may reflect introgression due to captive strains escaping into the wild. To identify specific genomic regions implicated, we used $f_{dM}$ to scan the genome, revealing genome wide introgression signals peaking over chromosome one (Figure 5B). Of the introgressed genetic material into the wild, 93% of the genes in this 2.1 Mb region were identified as Cytochrome P450s (supplementary table 9). Gene trees group introgressed wild individuals within a Sheppard-derived captive clade, corroborating the hypothesis of introgression (Figure 5C). Within lineages, we found evidence for introgression hotspots between lineage-*b*'s wild Europe and captive Australia individuals from multiple widespread locations, likely a product of anthropogenic mediated dispersal (Figure 5A). We also found evidence for bi-directional gene flow between wild western Europe and Australia indicative of repeated exchange between these geographically distant locations. This evidence supports similar observations of Black Soldier Fly colonisation of western Europe via admixed populations including Australasian ancestors [18].

We also detected introgression ($f_b$ = 0.152) into industry-strain-A (Ind-A) from wild Asian populations (Figure 5A). Genome-wide scans of $f_{dM}$ in Ind-A suggests a primarily lineage-*a* genetic background, indicated with a negative genome-wide mean $f_{dM}$, with a large 7.07 Mb region (S3: 150867943-157941741) of wild Asian origin introgressed on chromosome three (supplementary figure 13), containing a total of 121 genes (supplementary table 10). Several genes across this region have multiple copies including Collagen alpha (*Col4a1*; n=2), Dual oxidase (*Duox*; n=2) and salivary secreted peptide (*Sgsf*; n=5) genes (supplementary table 10).

## Discussion

Whole-genome sequencing of previously identified genetic clusters from populations of the Black Soldier Fly (*Hermetia illucens*) has provided insight into the natural diversity of this agriculturally important species. We have identified patterns associated with the genomic footprint of domestication. We find domesticated populations of lineage-*a* are largely derived from a North American wild population whereas lineage-*b* captive individuals appear to be naturalised within geographic regions likely due to their recent captivity or ongoing gene flow with the wild. Genetic diversity within this species is high and there are two highly divergent nuclear lineages. One of these lineages (-*b*) presents significant opportunity for establishing breeding programs due to its high genetic diversity in contrast to lineage-*a*, the lineage currently in widespread use by industry. Despite a worldwide distribution we observe an overall lack of hybridisation between these diverse lineages apart from a few examples of localised gene flow. Nonetheless, this may present opportunity for breeding strategies introducing novel wild genetic diversity into captive populations to improve performance [28]. The deep divergence between lineages may pose challenges for interbreeding programs due to hybrid breakdown, reduced fitness or pre-mating isolation, although recent crossing experiments suggest little evidence for incompatibilities [29].

**Revealing cryptic genetic diversity**

Whole-genome analysis reveals deep structure within the Black Soldier Fly. Previous mtDNA evidence estimated divergence between major clades over 2 my ago [20], but these data indicate even deeper divergences up to 6.8 – 8.4 my ago, greater than the divergence between *Drosophila* species [30–37]. Combined with a lack of evidence for extensive admixture, our data would suggest that Black Soldier Fly lineages could represent distinct species. Nonetheless, the genetic evidence indicates ongoing hybridisation suggesting a lack of complete reproductive incompatibilities [18,19], consistent with results from laboratory crosses [29]. Further behavioural and morphological analysis would be required before these lineages can be formally described as species or breed logs can be defined.

The existence of cryptic lineages could have important implications for the industry [38–40]. Progress in breeding programs could be hindered if incompatibilities and/or pre-mating isolation exist between strains [41,42]. Cryptic lineages within the Black Soldier Fly are likely in part responsible for the high phenotypic variance previously described [43], including nutritional [44–46] and performance variation [47,48]. Previous work has demonstrated significant phenotypic variation among strains of Sheppard and Australasian origin [49]. Within lineage-*a,* different populations and their interbreeding likely also contributed to this phenotypic variation reported across various studies. Among these samples, lineage-*b* is more diverse and might therefore hold a comparatively larger pool of unrecognised phenotypic variation of possible commercial relevance. Whereas at present most industrial activity harnesses lineage-*a* [18], emerging evidence suggests that identifying and optimising strain-substrate combinations could significantly improve performance, highlighting the importance of further research into evolutionary divergence and phenotypic plasticity of this species in the case of industrial application [49,50]. Overall, it remains unclear the extent to which commercially motivated worldwide exchange will contribute to the breakdown of these divergent lineages. However, deciphering the evolutionary history of the Black Soldier Fly will contribute to improving breeding and targeted genomic selection programs.

Our data also provides insights into the biogeographic history of this species. Whilst the within-species lineage divergence time is suggested to be older in the whole-genome data (3.9 – 13.8 my) unknown mutation rates and demographic history, including the differences in generation time between domesticated and wild populations may have inflated this prediction. Nonetheless, both mtDNA and whole-genome data in this and previous studies all show deep divergence within the Black Soldier Fly [20]. The deep divergences observed (>3 my ago) predate recent anthropogenic expansion, suggesting that this must represent divergence originating in the original native range in the Americas [17,18,20]. This could be a result of isolation-by-distance (IBD) across a broad range from South to North

America. Alternatively, deep divergence in the mtDNA and whole genome trees might reflect biogeographic events within the Americas. For example, divergence in mtDNA that occurred during the late Miocene period (~6 mya) coincides with uplift of the Andean Plateau, potentially driving reproductive barriers between South American populations [19,51–53]. The split between mtDNA lineages-*B1* and -*B2* occurred more recently (~3.2 my) which may represent colonisation of Central America after the rise of the isthmus of Panama [54]. Alternatively, the pattern of mito-nuclear discordance would be consistent with an introgression event over 3 million years ago whereby mtDNA lineage--*B1* introgressed into nuclear lineage-*a* [55]. Ancient hybridisation has been shown to facilitate adaptive ecological radiation and speciation which may have played a role in the diversification of this species [56]. Likewise, hybridisation may have played an important role in shaping the genetic basis for domestication and consequently, an entire industry largely building on lineage-*a* descendants. *Hermetia* are distributed throughout the globe, yet ~65% of species are found in the Neotropics [57], and only the Black Soldier Fly (*H. illucens*) is commonly studied. Therefore, *Hermetia* species are an interesting system to explore the consequences of hybridisation on geographic and ecological diversification. However, divergence time estimates should be interpreted with caution due to the absence of a Black Soldier Fly mutation rate. The mutation rate was inferred from *Drosophila*, which split from *Hermetia* ~200 mya. Variation in generation time and species-specific molecular mechanisms may have introduced significant error in these approximations [58].

**Genetic impacts of domestication**

Black Soldier Fly domestication has led to a loss of genetic diversity (39%) approaching that of domesticated crops (47-72%) [59,60]. This is perhaps surprising considering the recent timescale, estimated between one and two decades, of Black Soldier Fly domestication. Continued domestication will likely exacerbate this reduction of genetic diversity in commercial facilities raising further concerns around inbreeding and disease susceptibility [61,62]. Our whole-genome nuclear analysis identifies North America as the likely progenitor

of the common domesticated Black Soldier Fly Sheppard strain. Using published COI markers combined with this data further supports a progenitor of domesticated Black Soldier Fly to a wild North American origin [18–20]. This likely represents the single domestication event previously documented which established a colony over 200 generations ago [18]. In addition, we find evidence for at least a second independent domestication event, maybe more, occurring in captive populations in South America, Africa and Australia. Considerable genetic diversity and varying signals of domestication are present within described strains, potentially indicative of specific breeds. Whilst characteristic traits associated with distinct strains/ breeds have begun to be revealed, they remain largely unknown [43,49,50].

Unsurprisingly, over two decades of domestication occurring in parallel across the globe has shaped the genomes of farmed populations. Some caution should be applied when interpreting between signals of selection or drift, particularly when regions stretching over extreme distances are found, as inbreeding can also lead to runs of homozygosity and signatures similar to selective sweeps. Furthermore, our sample size for all populations apart from EVE were too small to reliably detect selective sweeps and therefore was not performed. Therefore, we analysed patterns of homozygosity associated with inbreeding, drift and selection and identified regions indicative of non-random genomic processes at play. Nonetheless, evidence suggests genomes of captive populations have been shaped by domestication and at least some of these regions may be signatures of selection. Although we cannot fully disentangle the contributions of selection and drift further research is required to disentangle these signatures with greater sample sizes. Despite this, domestication has played a significant role in shaping the genomic landscape of the Black Soldier Fly.

Domestication of the Black Soldier Fly may not only drive positive selection but, in the presence of abundant feed availability and optimal rearing conditions, may also lead to relaxed selection. Due to the nature of the industry, in which core breeding colonies are

typically reared on nutrient rich substrates, it is possible that domesticated strains could lose the ability to effectively feed on nutrient poor organic waste substrates, threatening efficient Feed Conversion Ratios (FCRs). This may reposition these lines useful for insect waste management over nutrient upcycling. On the other hand, increased rearing densities for commercial efficiency may also drive positive selection for stress tolerance and disease susceptibility [63]. The use of whole-genome sequencing data has provided high resolution insights into the impact of domestication on the genome of the Black Soldier Fly, previously undetected by microsatellites and mtDNA. Although it is challenging to differentiate drift and selection, it is important to further identify candidate genomic regions that could be explored further to understand domestication.

**Evidence of introgression and genomic consequences of gene flow**

Overall, there was limited evidence for introgression between major lineages of the Black Soldier Fly. In part, this may represent a bias in sampling as individuals representing major genetic clusters in the microsatellite data were chosen for sequencing here [18]. Despite this sampling, evidence of hybridisation was identified between lineages occurring in a small number of individuals in Europe, likely as a product of escaped flies from farms [20]. Genetic material found introgressed into a wild southern European individual on chromosome one showed an overrepresentation of P450 genes. This locus perhaps has an adaptive benefit in response to high mycotoxin concentration exposure in industrial organic waste streams [64,65]. However, these introgressed regions must undergo phenotypic validation to confirm their adaptive role. Whilst a lack of biosecurity measures allowing the release of domesticated flies into the wild poses conservation concerns, it also threatens potential undocumented and economically relevant trait loss from naturally diverse wild populations through genomic homogenisation [66,67]. In addition to genetic concerns, such uncontrolled releases may have broader ecological consequences. Domesticated populations may harbour unique disease threats to wild locally adapted populations [68].

There was also some limited evidence for gene flow into active commercial lines belonging to lineage-*a* from wild populations. Wild populations are notoriously difficult to establish in the lab but might harbour valuable phenotypic diversity that could be harnessed by developing introgressed lines through controlled breeding [61,69]. For example, we identified an introgressed genomic region derived from a wild Asian source within one industry population. All industry strains (A, B & C), sourced from Asian suppliers, showed evidence for admixture between lineage-*a* derived domesticated strains and wild Asian populations. Previous work has shown that admixed lineages are prevalent in production facilities, particularly across Asia [18]. Within an industrial setting introgression may promote genetic diversity [70] and enable the introduction of specific genes for novel applications [71]. Whilst extensive introgression may protect against colony collapse in the short term, establishing an industry with a single founding stock population with a highly similar genetic background severely increases the risk of disease susceptibility in the long term [61,62,72,73]. Clearly further work is needed to understand the degree of heterosis versus hybrid breakdown in crosses between divergent lineages [28]. In the light of this genetic diversity, cost effective and accessible genetic management will likely play an important role in this industry in the future.

## Conclusions

Our study indicates the existence of previously undescribed cryptic diversity within the Black Soldier Fly. Levels of divergence observed are in the same order of magnitude as divergent *Drosophila* species. This suggests a potential need for a taxonomic revaluation or breed book record keeping of the Black Soldier Fly following further sampling efforts in the Americas. We demonstrate that two decades of extensive captive rearing, originating from a probable wild North American progenitor, has led to dramatic changes in the genome. Importantly, irrespective of their distribution, captive rearing has left distinct signatures associated with domestication across genetically distinct populations. This species is an

excellent system to study evolution due to the rapid and ongoing domestication pressures on an industrial and global scale. The genomic diversity described here sets the stage for future research to better characterise the genetic identity of strains and to monitor inbreeding in an industrial setting. Further sampling will be needed to elucidate the genetic basis of domestication, but this species could become an interesting case study for understanding human impacts on domesticated species as well as providing a foundation for genetic improvements.

## Materials and Methods

### Sample collection, whole genome sequencing and variant calling

We obtained 54 whole-bodied (adult) Black Soldier Fly and one *Ptecticus aurifer* sample for whole genome sequencing (WGS). Individuals were obtained from either previously sampled collections [18], maintained (but sourced from three independent Asian suppliers) within industry conditions (Better Origin, UK) or from a taxonomic voucher set (source: M. Hauser). Total genomic DNA was extracted from thorax muscle tissue and purified using the Qiagen Blood & Culture DNA Midi Kit (Qiagen, Germany). Libraries ($n$ = 39) were prepared, and genome sequenced by BGI (BGI, Hong Kong) as in Generalovic et al. (2021), including the 12 previously published genomes. Novogene (Cambridge, UK) constructed a second set of libraries ($n$ = 11) and performed paired end (PE) 150 bp sequencing on the NovaSeq platform. The remaining five libraries were sequenced in the same manner but constructed using a custom Tn5 based protocol [74]. Raw sequencing data was quality checked, with FastQC v0.11.4 [75] and visualised in MultiQC v1.9 [76] both pre- and post- adapter trimming with cutadapt v1.8.1 [77]. Treated reads were mapped to the GCA_905115235.1 reference genome [25] using bwa-mem v0.7.17 algorithm [78]. Resulting BAM files were sorted, duplicates marked and removed using samtools v1.9 [79] and picard v2.9.2 (http://broadinstitute.github.io/picard). Variant calling was performed following the best practices of Genome Analysis Toolkit (GATK) v3.7.0 [80]. Genotyped SNPs were quality

checked and filtered for high quality (>75) sites setting a minimum and maximum depth cut-off (10-34X) and ignoring sites containing a high proportion of missingness (>15%). Assessment of contaminants based on allelic balance within samples was carried out using allelicBalance.py (https://github.com/joanam/scripts/blob/master/allelicBalance.py). An additional VCF dataset was generated using a minor allele frequency (MAF) filter threshold of 0.1.

**Mitochondrial *de novo* assembly**

Analysis of mitochondrial genomes was performed by assembling *de novo* whole mitochondrial genomes using a subset of the novel WGS reads produced in this study and collected from previous work [20,25,26]. We assumed, due to high mitochondria copy number in the muscle tissue used for DNA extraction, that mitochondrial DNA was enriched in our WGS data. We used a minimum subset of 1.6 Gb raw PE data as input for *de novo* assembly per individual and a maximum of 4.8 Gb. Mitochondrial *de novo* assembly was carried out using the MitoZ toolkit v2.4 [81]. We used the 'all' function to assemble, annotate and plot genomes with parameters 'genetic_code=5' and 'clade=Arthropoda'. In cases where assemblies yielded the reverse compliment of the mitogenome, identified by the gene order and orientation being the opposite to that of the ancestral insect [82], the mitogenome was derived by running the assembly through the reverse compliment function of MEGA-X v10.2.6 [83]. An additional 57 mitochondrial genomes were obtained from previous studies [20,25,26].

**Phylogenetic and population structure**

A Maximum Likelihood (ML) phylogenetic tree was constructed using whole nuclear genomes to assess phylogenetic structure and relationships within the Black Soldier Fly species. We used RAxML program v8.2.9 [84] to construct a ML tree using the generalized time-reversible substitution model and gamma (GTRGAMMA) model, Lewis ascertainment

bias and 100 bootstrap (BS) replications based on a random subset of 648,059 nuclear whole-genome SNPs. Further phylogenetic analysis using ASTRAL v5.7.8 [85], a coalescent based method robust to Incomplete Lineage Sorting (ILS), was performed to validate the ML analysis. A 10% sample of SNPs from each chromosome was extracted and split into 5,000 SNP segments and a gene tree was constructed on each using 100 BS replications totalling 8,800 trees, this was repeated for both the Neighbour-Joining (NJ) and ML method.

Phylogenetic reconstruction of the mitochondrial genomes was also performed. Whole mitochondrial genomes were aligned using MUSCLE v3.8.31 [86] and a phylogenetic ML tree produced using RAxML v8.2.9 [83]. A principal component analysis (PCA) [87] was performed using plink v1.9 [88] with Black Soldier Fly samples only. A random subsample of two individuals per unique geographic sampled population was used to conduct the PCA analysis after it was found that an unbalanced design artificially inflated genetic variation across both PC1 and PC2. Structure analysis was performed using ADMIXTURE v1.3.0 [89] assuming a minimum and maximum number of ancestral populations (*k*) of two and 11, respectively, based on the number of populations sampled. Linkage pruning was performed on SNPs used for both PCA and STRUCTURE analysis using plink v1.9 (--double-id --allow-extra-chr --set-missing-var-ids @:# --indep-pairwise 50 10 0.1) [88]. This filtering reduced the impact of non-independent SNPs reducing bias when estimating population structure ensuring sites were not disproportionately driven by physically linked loci. A NJ tree [90] was constructed using wild and captive Cytochrome Oxidase I (COI) mitochondrial marker genes from published and *de novo* assembled mitochondrial genomes. A NJ analysis of COI sequences was performed by aligning sequences with MUSCLE v3.8.31 [86] and phylogenetic analysis with MEGA v10.1.7 [83] using default parameters.

**Intraspecific divergence, diversity and genetic differentiation**

Patterns of intraspecific divergence within the Black Soldier Fly complex were revealed using genome wide (nuclear) scans for genetic differentiation ($F_{ST}$) and sequence divergence ($D_{XY}$)

on all populations. Non-overlapping 20 kb windows in 20 kb steps containing a minimum of 100 genotyped sites (-w 20000 -m 100 -s 20000) were used to generate genome-wide Nucleotide diversity ($\pi$), $F_{ST}$ and $D_{XY}$ statistics using popgenWindows.py (https://github.com/simonhmartin/genomics_general). Genome-wide nucleotide diversity was plotted in 20 kb windows using data smoothing (method = "loess", span = 0.01).

**Divergence time analysis**

Due to the paucity of Black Soldier Fly divergence data, a dated Diptera tree was first required to provide temporal calibration points for a Black Soldier Fly divergence analysis, as in Nardi *et al.* (2010) [91]. Here, we used mitochondrial DNA (mtDNA) only to estimate approximate divergences times. For the first calibration dating run mitogenomes of two Black Soldier Fly individuals, from divergent branches of both the nuclear and mitochondrial phylogeny and a single outgroup individual, *Ptecticus aurifer,* were obtained along with ten Diptera mitogenomes from GenBank (supplementary table 11) [92]. These dipteran mitochondrial genomes were aligned in MUSCLE v3.8.31 [86]. A molecular clock test was conducted on the resultant alignment and the optimal model of sequence evolution was selected by AIC using MEGA-X v10.2.6 [83]. A Bayesian dating analysis was conducted in BEAST v2.6.6 [93], using a GTR+G+I model of sequence evolution, and a relaxed lognormal molecular clock. Dating priors were set to calibrate six nodes spanning the Diptera phylogeny, based on fossil and biogeographic evidence, and the Yule model was used for the tree prior (supplementary table 4). Markov chain Monte Carlo (MCMC) length was set to 100 million, with samples logged every 1,000 steps. The model output was inspected in Tracer v1.7.2 [94]. After checking for parameter convergence, the burn-in was set to 25%, and the effective sample size of relevant parameters was verified to exceed 150.

A second BEAST run was conducted using temporal constraints derived in the first run for the Stratiomyidae and *H.illucens* nodes. This allowed estimation of key divergences within the Black Soldier Fly phylogeny, given the lack of molecular dating studies, individuals

which spanned the phylogeny were used (supplementary table 4), alongside a *P. aurifer* outgroup. Alignment and divergence analysis was repeated a second time using the parameters previously described. We used a tree prior with a coalescent model of constant population size. Temporal constraints were placed on two nodes, using the output of the initial BEAST run, corresponding to the *P. aurifer* - Black Soldier Fly split and the common ancestor of the Black Soldier Fly clade (supplementary table 4). The MCMC chain was as previously described. After inspecting the model output in Tracer v1.7.2 [94], a burn-in of 10% was set. The dated phylogeny was derived using TreeAnnotator package of BEAST v2.6.6 [93]. The mitochondrial sequences used for both BEAST runs and their GenBank accession numbers are given in supplementary table 11, while all temporal constraints, outputs, and literature corroboration for phylogenetic nodes spanning both dating runs are given in supplementary table 4.

To corroborate mtDNA divergence time results we also estimated divergence times of Black Soldier Fly lineages using whole genome data. A random subsample of 3,978 genome-wide SNPs were used from all sequenced individuals for Bayesian dating analysis, again conducted in BEAST v2.7.7 [93]. Sex chromosome S7 was excluded from this analysis. A single temporal constraint was set for the *P. aurifer* and *H. illucens* split as our previous evidence and published evidence [95] were in agreement. A coalescent constant population model was used with the *Drosophila melanogaster* mutation rate of $2.8 \times 10^{-9}$ as an estimate for the Black Soldier Fly mutation rate [96]. A starting tree was provided based on the phylogenetic tree produced in Figure 1A. A chain length of 10 million was used with a burn-in rate of 10% and sampling was performed every 1,000 chains. Our effective sample size was deemed suitable as it exceeded 4,000. Results were assessed and visualised in Tracer v1.7.2 [94].

**Genetic signals of domestication**

Genomic diversity, and Linkage Disequilibrium (LD) decay were calculated to reveal genetic signatures of domestication in sampled populations with a sequencing depth >19x.

Nucleotide diversity ($\pi$) and $F_{ST}$ was calculated in non-overlapping 100 kb windows using popgenWindows.py (-w 100000 -m 100 -s 100000) (https://github.com/simonhmartin/genomics_general) using 100 kb steps containing at least 100 genotyped sites. Values of $\pi$ for every individual was presented as mean heterozygosity within each sampled population. Patterns of LD decay were first calculated genome wide for a random subset of three individuals for each population. We used PopLDdecay v3.41 [97] to generate pairwise LD $r^2$ values setting a maximum distance of 50 kb between sampled SNPs.

To assess the impact of domestication in the Black Soldier Fly we performed scans for Runs of Homozygosity (RoH) to reveal genomic regions associated with positive selection and genetic drift. We tested each maintained captive population independently despite known shared population history and strong *k*-means clustering support for a single population structure of the Sheppard-lineage, due to a documented split approximately 1-2 decades ago. We also used patterns observed from *k*-means ancestry assignment (admixture) analysis to support our grouping into geographic clusters . All geographic populations were treated independently based off results at *k* = 11. Industry strains (A-C) were grouped based on population structure results and identical rearing conditions. Runs of Homozygosity were identified as in Generalovic *et al*. (2021) using the MAF (0.1) filtered VCF file with the following exceptions “windowSize = 20”, “minSNP = 20” and “minLengthBps = 250000” using the detectRUNS v0.98 R package [98]. Chromosome seven (X chromosome) and individuals with no detected homozygous regions were excluded from RoH analysis. Additional support using $F_{ST}$ with the closest sampled wild progenitor was acquired performed as in “Genetic signals of domestication”. Candidate loci were identified from gene codes in the GCA_905115235.1 reference by performing blastp searches against the Arthropoda blast nr database using an *e*-value threshold of 1*e*-5. Gene Ontology summary information for each candidate gene was obtained from flybase [99].

**Introgression and gene flow**

Mito-nuclear discordance was predicted by plotting a co-phylogenetic plot of the previously constructed nuclear and mitochondrial phylogenies using the phytools v1.0-3 package [100] in R v3.6.3 [101]. Patterns of admixture were detected by revealing excess allele sharing between individuals using ancestral (ABBA) and derived (BABA) allele patterns [102,103], $f_{dM}$ values [104] and the *f*-branch statistic [105]. We used ABBA/BABA patterns using Dsuite v0.4r38 Dtrios function [106] and provided the previously resolved SNP based phylogeny (Figure 1A) as a tree file (-t). Trios showing *p*-values > 0.05 and Z-scores < 3 post *p*-value adjustment using Benjamini-Hochberg multiple correction in R v4.0.3 [102] were filtered out of the analysis. We generated the *f*-branch ($f_b$) statistic using individuals sequenced to the highest coverage to represent populations within the phylogeny to assign gene flow to all, including internal ancestral, branches using *P. aurifer* as the outgroup [105]. Populations that were not monophyletic were represented by multiple individuals in the $f_b$ analysis. Genome-wide scans were performed on samples showing an excess of gene flow identified from the *f*-branch method using the Dsuite Dinvestigate function calculating average $f_{dM}$ over 50 SNPs in 25 SNP steps across the whole genome. Genomic hotspots of introgression were identified by setting an arbitrary 1% cut-off for $f_{dM}$ values.

## Acknowledgments

This work was supported by the Biotechnology and Biological Sciences Research Council (BB/M011194/1).

## Availability of supporting data

The data sets supporting the results of this article are available under Bioprojects PRJEB37575 and PRJEB58720.

## References

1. Kareiva P, Watts S, McDonald R, Boucher T. Domesticated Nature: Shaping Landscapes and Ecosystems for Human Welfare. Science. 2007;316: 1866–1869. doi:10.1126/science.1140170

2. Purugganan MD. What is domestication? Trends in Ecology & Evolution. 2022;37: 663–671. doi:10.1016/j.tree.2022.04.006

3. Zeder MA. Core questions in domestication research. Proc Natl Acad Sci USA. 2015;112: 3191–3198. doi:10.1073/pnas.1501711112

4. Mabry ME, Turner-Hissong SD, Gallagher EY, McAlvay AC, An H, Edger PP, et al. The Evolutionary History of Wild, Domesticated, and Feral *Brassica oleracea* (Brassicaceae). Molecular Biology and Evolution. 2021;38: 4419–4434. doi:10.1093/molbev/msab183

5. Ostrander EA, Wayne RK, Freedman AH, Davis BW. Demographic history, selection and functional diversity of the canine genome. Nat Rev Genet. 2017;18: 705–720. doi:10.1038/nrg.2017.67

6. Darwin C. The variation of animals and plants under domestication. Daya Books; 1885.

7. Larson G, Piperno DR, Allaby RG, Purugganan MD, Andersson L, Arroyo-Kalin M, et al. Current perspectives and the future of domestication studies. Proc Natl Acad Sci USA. 2014;111: 6139–6146. doi:10.1073/pnas.1323964111

8. Wang M-S, Thakur M, Peng M-S, Jiang Y, Frantz LAF, Li M, et al. 863 genomes reveal the origin and domestication of chicken. Cell Res. 2020;30: 693–701. doi:10.1038/s41422-020-0349-y

9. Kwon YM, Vranken N, Hoge C, Lichak MR, Norovich AL, Francis KX, et al. Genomic consequences of domestication of the Siamese fighting fish. Sci Adv. 2022;8: eabm4950. doi:10.1126/sciadv.abm4950

10. Lecocq T. Insects: the disregarded domestication histories. Animal domestication. IntechOpen; 2018.

11. Tomberlin JK, Van Huis A. Black soldier fly from pest to ‘crown jewel’ of the insects as feed industry: an historical perspective. JIFF. 2020;6: 1–4. doi:10.3920/JIFF2020.0003

12. Sheppard DC, Tomberlin JK, Joyce JA, Kiser BC, Sumner SM. Rearing Methods for the Black Soldier Fly (Diptera: Stratiomyidae). J Med Entomol. 2002;39: 695–698. doi:10.1603/0022-2585-39.4.695

13. Van Huis A. Potential of Insects as Food and Feed in Assuring Food Security. Annu Rev Entomol. 2013;58: 563–583. doi:10.1146/annurev-ento-120811-153704

14. Van Huis A, Van Itterbeeck J, Klunder H, Mertens E, Halloran A, Muir G, et al. Edible insects: future prospects for food and feed security. 2013.

15. Surendra KC, Tomberlin JK, Van Huis A, Cammack JA, Heckmann L-HL, Khanal SK. Rethinking organic wastes bioconversion: Evaluating the potential of the black soldier fly (*Hermetia illucens* (L.)) (Diptera: Stratiomyidae) (BSF). Waste Management. 2020;117: 58–80. doi:10.1016/j.wasman.2020.07.050

16. Gold M, Tomberlin JK, Diener S, Zurbrügg C, Mathys A. Decomposition of biowaste macronutrients, microbes, and chemicals in black soldier fly larval treatment: A review. Waste Management. 2018;82: 302–318. doi:10.1016/j.wasman.2018.10.022

17. Marshall S, Woodley N, Hauser M. The historical spread of the black soldier fly, *Hermetia illucens* (L.)(Diptera, Stratiomyidae, Hermetiinae), and its establishment in Canada. The Journal of the Entomological Society of Ontario. 2015;146.

18. Kaya C, Generalovic TN, Ståhls G, Hauser M, Samayoa AC, Nunes-Silva CG, et al. Global population genetic structure and demographic trajectories of the black soldier fly, *Hermetia illucens*. BMC Biol. 2021;19: 94. doi:10.1186/s12915-021-01029-w

19. Ståhls G, Meier R, Sandrock C, Hauser M, Šašić Zorić L, Laiho E, et al. The puzzling mitochondrial phylogeography of the black soldier fly (*Hermetia illucens*), the commercially most important insect protein species. BMC Evol Biol. 2020;20: 60. doi:10.1186/s12862-020-01627-2

20. Guilliet J, Baudouin G, Pollet N, Filée J. What complete mitochondrial genomes tell us about the evolutionary history of the black soldier fly, *Hermetia illucens*. BMC Ecol Evo. 2022;22: 72. doi:10.1186/s12862-022-02025-6

21. Facchini E, Shrestha K, Van Den Boer E, Junes P, Sader G, Peeters K, et al. Long-Term Artificial Selection for Increased Larval Body Weight of *Hermetia illucens* in Industrial Settings. Front Genet. 2022;13: 865490. doi:10.3389/fgene.2022.865490

22. Generalovic TN, Zhou W, Zhao LC, Leonard S, Warren IA, Pipan M, et al. Repeatable phenotypic but not genetic response to selection on body size in the black soldier fly. Evolutionary Biology; 2025. doi:10.1101/2025.02.25.640052

23. Zimmerman SJ, Aldridge CL, Oyler-McCance SJ. An empirical comparison of population genetic analyses using microsatellite and SNP data for a species of conservation concern. BMC Genomics. 2020;21: 382. doi:10.1186/s12864-020-06783-9

24. Fischer MC, Rellstab C, Leuzinger M, Roumet M, Gugerli F, Shimizu KK, et al. Estimating genomic diversity and population differentiation – an empirical comparison of microsatellite and SNP variation in *Arabidopsis halleri*. BMC Genomics. 2017;18: 69. doi:10.1186/s12864-016-3459-7

25. Generalovic TN, McCarthy SA, Warren IA, Wood JMD, Torrance J, Sims Y, et al. A high-quality, chromosome-level genome assembly of the Black Soldier Fly (*Hermetia illucens* L.). G3 Genes|Genomes|Genetics. 2021;11: jkab085. doi:10.1093/g3journal/jkab085

26. Qi Y, Xu J, Tian X, Bai Y, Gu X. The complete mitochondrial genome of *Hermetia illucens* (Diptera: Stratiomyidae). Mitochondrial DNA Part B. 2017;2: 189–190. doi:10.1080/23802359.2017.1307708

27. Zhan Q, Zhao Y, Zhang S, Peng X. The complete mitochondrial genome of the soldier fly *Ptecticus aurifer*. Mitochondrial DNA Part B. 2020;5: 660–661. doi:10.1080/23802359.2020.1711820

28. Lippman ZB, Zamir D. Heterosis: revisiting the magic. Trends in Genetics. 2007;23: 60–66. doi:10.1016/j.tig.2006.12.006

29. Meyermans R, Broeckx L, Mondelaers J, Gorssen W, Frooninckx L, Janssens S, Van Miert S, Buys N. Exploring the potential of crossbreeding to enhance black soldier fly (*Hermetia illucens*) production. Journal of Insects as Food and Feed, *11*(10), 1703-1715. doi:10.1163/23524588-00001391

30. Irvin SD, Wetterstrand KA, Hutter CM, Aquadro CF. Genetic Variation and Differentiation at Microsatellite Loci in *Drosophila simulans*: Evidence for Founder Effects in New World Populations. Genetics. 1998;150: 777–790. doi:10.1093/genetics/150.2.777

31. Caracristi G. Genetic Differentiation Between American and European *Drosophila melanogaster* Populations Could Be Attributed to Admixture of African Alleles. Molecular Biology and Evolution. 2003;20: 792–799. doi:10.1093/molbev/msg091

32. Verspoor RL, Haddrill PR. Genetic Diversity, Population Structure and Wolbachia Infection Status in a Worldwide Sample of *Drosophila melanogaster* and *D. simulans* Populations. PLoS ONE. 2011;6: e26318. doi:10.1371/journal.pone.0026318

33. Nolte V, Pandey RV, Kofler R, Schlötterer C. Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana*. Genome Res. 2013;23: 99–110. doi:10.1101/gr.139873.112

34. Kao JY, Zubair A, Salomon MP, Nuzhdin SV, Campo D. Population genomic analysis uncovers African and European admixture in *Drosophila melanogaster* populations from the south-eastern United States and Caribbean Islands. Molecular Ecology. 2015;24: 1499–1509. doi:10.1111/mec.13137

35. Meiklejohn CD, Landeen EL, Gordon KE, Rzatkiewicz T, Kingan SB, Geneva AJ, et al. Gene flow mediates the role of sex chromosome meiotic drive during complex speciation. Evolutionary Biology; 2015. doi:10.1101/024711

36. Kang L, Rashkovetsky E, Michalak K, Garner HR, Mahaney JE, Rzigalinski BA, et al. Genomic divergence and adaptive convergence in *Drosophila simulans* from Evolution Canyon, Israel. Proc Natl Acad Sci USA. 2019;116: 11839–11844. doi:10.1073/pnas.1720938116

37. Tamura K. Temporal Patterns of Fruit Fly (Drosophila) Evolution Revealed by Mutation Clocks. Molecular Biology and Evolution. 2003;21: 36–44. doi:10.1093/molbev/msg236

38. Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, et al. Cryptic species as a window on diversity and conservation. Trends in Ecology & Evolution. 2007;22: 148–155. doi:10.1016/j.tree.2006.11.004

39. Ghai RR, Chapman CA, Omeja PA, Davies TJ, Goldberg TL. Nodule Worm Infection in Humans and Wild Primates in Uganda: Cryptic Species in a Newly Identified Region of Human Transmission. PLoS Negl Trop Dis. 2014;8: e2641. doi:10.1371/journal.pntd.0002641

40. Siddall ME, Trontelj P, Utevsky SY, Nkamany M, Macdonald KS. Diverse molecular data demonstrate that commercially available medicinal leeches are not *Hirudo medicinalis*. Proc R Soc B. 2007;274: 1481–1487. doi:10.1098/rspb.2007.0248

41. Mérot C, Salazar C, Merrill RM, Jiggins CD, Joron M. What shapes the continuum of reproductive isolation? Lessons from *Heliconius* butterflies. Proc R Soc B. 2017;284: 20170335. doi:10.1098/rspb.2017.0335

42. Maltseva AL, Varfolomeeva MA, Lobov AA, Tikanova PO, Repkin EA, Babkina IY, et al. Premating barriers in young sympatric snail species. Scientific Reports. 2021;11: 5720.

43. Zhou F, Tomberlin JK, Zheng L, Yu Z, Zhang J. Developmental and Waste Reduction Plasticity of Three Black Soldier Fly Strains (Diptera: Stratiomyidae) Raised on Different Livestock Manures. jnl med entom. 2013;50: 1224–1230. doi:10.1603/ME13021

44. Spranghers T, Ottoboni M, Klootwijk C, Ovyn A, Deboosere S, De Meulenaer B, et al. Nutritional composition of black soldier fly (*Hermetia illucens*) prepupae reared on different organic waste substrates. J Sci Food Agric. 2017;97: 2594–2600. doi:10.1002/jsfa.8081

45. Meneguz M, Schiavone A, Gai F, Dama A, Lussiana C, Renna M, et al. Effect of rearing substrate on growth performance, waste reduction efficiency and chemical composition of black soldier fly (*Hermetia illucens*) larvae. J Sci Food Agric. 2018;98: 5776–5784. doi:10.1002/jsfa.9127

46. Liland NS, Biancarosa I, Araujo P, Biemans D, Bruckner CG, Waagbø R, et al. Modulation of nutrient composition of black soldier fly (*Hermetia illucens*) larvae by feeding seaweed-enriched media. PLoS ONE. 2017;12: e0183188. doi:10.1371/journal.pone.0183188

47. Rodrigues DP, Calado R, Pinho M, Rosário Domingues M, Antonio Vázquez J, Ameixa OMCC. Bioconversion and performance of Black Soldier Fly (*Hermetia illucens*) in the recovery of nutrients from expired fish feeds. Waste Management. 2022;141: 183–193. doi:10.1016/j.wasman.2022.01.035

48. Bosch G, Van Zanten HHE, Zamprogna A, Veenenbos M, Meijer NP, Van Der Fels-Klerx HJ, et al. Conversion of organic resources by black soldier fly larvae: Legislation, efficiency and environmental impact. Journal of Cleaner Production. 2019;222: 355–363. doi:10.1016/j.jclepro.2019.02.270

49. Sandrock C, Leupi S, Wohlfahrt J, Kaya C, Heuel M, Terranova M, et al. Genotype-by-Diet Interactions for Larval Performance and Body Composition Traits in the Black Soldier Fly, *Hermetia illucens*. Insects. 2022;13: 424. doi:10.3390/insects13050424

50. Generalovic TN, Sandrock C, Leonard S, Schuldiner-Harpaz T, Pipan M, Welch JJ, et al. Variation in strain performance and estimates of heritability of body size indicate considerable potential for genetic improvement of the black soldier fly (*Hermetia illucens*). Entomologia Exp Applicata. 2025;173: 558–574. doi:10.1111/eea.13565

51. Garzione CN, Molnar P, Libarkin JC, MacFadden BJ. Rapid late Miocene rise of the Bolivian Altiplano: Evidence for removal of mantle lithosphere. Earth and Planetary Science Letters. 2006;241: 543–556. doi:10.1016/j.epsl.2005.11.026

52. Ghosh P, Garzione CN, Eiler JM. Rapid Uplift of the Altiplano Revealed Through 13C-18O Bonds in Paleosol Carbonates. Science. 2006;311: 511–515. doi:10.1126/science.1119365

53. Pazmiño MF, Del Hierro AG, Flores FJ. Genetic diversity and organic waste degrading capacity of *Hermetia illucens* from the evergreen forest of the Equatorial Choco lowland. PeerJ. 2023;11: e14798. doi:10.7717/peerj.14798

54. O'Dea A, Lessios HA, Coates AG, Eytan RI, Restrepo-Moreno SA, Cione AL, et al. Formation of the Isthmus of Panama. Sci Adv. 2016;2: e1600883. doi:10.1126/sciadv.1600883

55. Excoffier L, Foll M, Petit RJ. Genetic Consequences of Range Expansions. Annu Rev Ecol Evol Syst. 2009;40: 481–501. doi:10.1146/annurev.ecolsys.39.110707.173414

56. Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. Ancient hybridization fuels rapid cichlid fish adaptive radiations. Nat Commun. 2017;8: 14363. doi:10.1038/ncomms14363

57. Fachin DA, Hauser M. Large Flies Overlooked: the Genus Hermetia Latreille, 1804 (Diptera: Stratiomyidae) in the Neotropics, with 11 Synonyms and a New Species to Brazil. Neotrop Entomol. 2022;51: 660–690. doi:10.1007/s13744-022-00978-3

58. Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim J-W, Lambkin C, et al. Episodic radiations in the fly tree of life. Proc Natl Acad Sci USA. 2011;108: 5690–5695. doi:10.1073/pnas.1012675108

59. Page A, Gibson J, Meyer RS, Chapman MA. Eggplant Domestication: Pervasive Gene Flow, Feralization, and Transcriptomic Divergence. Molecular Biology and Evolution. 2019;36: 1359–1372. doi:10.1093/molbev/msz062

60. Bitocchi E, Bellucci E, Giardini A, Rau D, Rodriguez M, Biagetti E, et al. Molecular analysis of the parallel domestication of the common bean (*P haseolus vulgaris*) in Mesoamerica and the Andes. New Phytologist. 2013;197: 300–313. doi:10.1111/j.1469-8137.2012.04377.x

61. Rhode C, Badenhorst R, Hull KL, Greenwood MP, Bester-van Der Merwe AE, Andere AA, et al. Genetic and phenotypic consequences of early domestication in black soldier flies (*Hermetia illucens*). Animal Genetics. 2020;51: 752–762. doi:10.1111/age.12961

62. Spielman D, Brook BW, Briscoe DA, Frankham R. Does Inbreeding and Loss of Genetic Diversity Decrease Disease Resistance? Conservation Genetics. 2004;5: 439–448. doi:10.1023/B:COGE.0000041030.76598.cd

63. Hull KL, Greenwood MP, Lloyd M, Bester-van Der Merwe AE, Rhode C. Gene expression differentials driven by mass rearing and artificial selection in black soldier fly colonies. Insect Molecular Biology. 2023;32: 86–105. doi:10.1111/imb.12816

64. Zhan S, Fang G, Cai M, Kou Z, Xu J, Cao Y, et al. Genomic landscape and genetic manipulation of the black soldier fly *Hermetia illucens*, a natural waste recycler. Cell Res. 2020;30: 50–60. doi:10.1038/s41422-019-0252-6

65. Meijer N, Stoopen G, Van Der Fels-Klerx HJ, Van Loon JJA, Carney J, Bosch G. Aflatoxin B1 Conversion by Black Soldier Fly (*Hermetia illucens*) Larval Enzyme Extracts. Toxins. 2019;11: 532. doi:10.3390/toxins11090532

66. McGinnity P, Prodöhl P, Ferguson A, Hynes R, Maoiléidigh NÓ, Baker N, et al. Fitness reduction and potential extinction of wild populations of Atlantic salmon, *Salmo salar* , as a result of interactions with escaped farm salmon. Proc R Soc Lond B. 2003;270: 2443–2450. doi:10.1098/rspb.2003.2520

67. Fraser DJ, Minto C, Calvert AM, Eddington JD, Hutchings JA. Potential for domesticated–wild interbreeding to induce maladaptive phenology across multiple populations of wild Atlantic salmon (*Salmo salar*). Taylor E, editor. Can J Fish Aquat Sci. 2010;67: 1768–1775. doi:10.1139/F10-094

68. Pienaar RD, Herrero S, Cerqueira De Araujo A, Krupa F, Abd-Alla AMM, Herniou EA. High-throughput screening reveals high diversity and widespread distribution of viruses in black soldier flies (*Hermetia illucens*). Journal of Invertebrate Pathology. 2025;211: 108322. doi:10.1016/j.jip.2025.108322

69. Kumar A, Sandhu N, Venkateshwarlu C, Priyadarshi R, Yadav S, Majumder RR, et al. Development of introgression lines in high yielding, semi-dwarf genetic backgrounds to enable improvement of modern rice varieties for tolerance to multiple abiotic stresses free from undesirable linkage drag. Sci Rep. 2020;10: 13073. doi:10.1038/s41598-020-70132-9

70. Hegstad JM, Nelson RL, Renny-Byfield S, Feng L, Chaky JM. Introgression of novel genetic diversity to improve soybean yield. Theor Appl Genet. 2019;132: 2541–2552. doi:10.1007/s00122-019-03369-2

71. Burgarella C, Barnaud A, Kane NA, Jankowski F, Scarcelli N, Billot C, et al. Adaptive Introgression: An Untapped Evolutionary Mechanism for Crop Adaptation. Front Plant Sci. 2019;10: 4. doi:10.3389/fpls.2019.00004

72. Bryden CA, Heath JW, Heath DD. Performance and heterosis in farmed and wild Chinook salmon (*Oncorhynchus tshawytscha*) hybrid and purebred crosses. Aquaculture. 2004;235: 249–261. doi:10.1016/j.aquaculture.2004.01.027

73. Doyle RW. Inbreeding and disease in tropical shrimp aquaculture: a reappraisal and caution. Aquac Res. 2016;47: 21–35. doi:10.1111/are.12472

74. Picelli S, Björklund ÅK, Reinius B, Sagasser S, Winberg G, Sandberg R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 2014;24: 2033–2040. doi:10.1101/gr.177881.114

75. Simon Andrews F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. FastQC: a quality control tool for high throughput sequence data. 2010.

76. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32: 3047–3048. doi:10.1093/bioinformatics/btw354

77. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. 2011;17: 10. doi:10.14806/ej.17.1.200

78. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25: 1754–1760. doi:10.1093/bioinformatics/btp324

79. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

80. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

81. Meng G, Li Y, Yang C, Liu S. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Research. 2019;47: e63–e63. doi:10.1093/nar/gkz173

82. Cameron SL. Insect Mitochondrial Genomics: Implications for Evolution and Phylogeny. Annu Rev Entomol. 2014;59: 95–117. doi:10.1146/annurev-ento-011613-162007

83. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Battistuzzi FU, editor. Molecular Biology and Evolution. 2018;35: 1547–1549. doi:10.1093/molbev/msy096

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

85. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 2018;19: 153. doi:10.1186/s12859-018-2129-y

86. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5: 113. doi:10.1186/1471-2105-5-113

87. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006;2: e190. doi:10.1371/journal.pgen.0020190

88. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007;81: 559–575. doi:10.1086/519795

89. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664. doi:10.1101/gr.094052.109

90. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987. doi:10.1093/oxfordjournals.molbev.a040454

91. Nardi F, Carapelli A, Boore JL, Roderick GK, Dallai R, Frati F. Domestication of olive fly through a multi-regional host shift to cultivated olives: Comparative dating using complete mitochondrial genomes. Molecular Phylogenetics and Evolution. 2010;57: 678–686. doi:10.1016/j.ympev.2010.08.008

92. NCBI Resource Coordinators, Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2018;46: D8–D13. doi:10.1093/nar/gkx1095

93. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. Pertea M, editor. PLoS Comput Biol. 2019;15: e1006650. doi:10.1371/journal.pcbi.1006650

94. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. Susko E, editor. Systematic Biology. 2018;67: 901–904. doi:10.1093/sysbio/syy032

95. Brammer CA, Von Dohlen CD. Evolutionary history of Stratiomyidae (Insecta: Diptera): The molecular phylogeny of a diverse family of flies. Molecular Phylogenetics and Evolution. 2007;43: 660–673. doi:10.1016/j.ympev.2006.09.006

96. Keightley PD, Ness RW, Halligan DL, Haddrill PR. Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a *Drosophila melanogaster* Full-Sib Family. Genetics. 2014;196: 313–320. doi:10.1534/genetics.113.158758

97. Zhang C, Dong S-S, Xu J-Y, He W-M, Yang T-L. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics. 2019;35: 1786–1788. doi:10.1093/bioinformatics/bty875

98. Biscarini F, Cozzi P, Gaspa G, Marras G, others. Detect runs of homozygosity and runs of heterozygosity in diploid genomes. 2019.

99. Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, et al. FlyBase 2.0: the next generation. Nucleic Acids Research. 2019;47: D759–D765. doi:10.1093/nar/gky1003

100. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012;3: 217–223. doi:10.1111/j.2041-210X.2011.00169.x

101. RStudio Team. RStudio: integrated development for R. 2015.

102. Durand EY, Patterson N, Reich D, Slatkin M. Testing for Ancient Admixture between Closely Related Populations. Molecular Biology and Evolution. 2011;28: 2239–2252. doi:10.1093/molbev/msr048

103. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010;328: 710–722. doi:10.1126/science.1188021

104. Malinsky M, Challis RJ, Tyers AM, Schiffels S, Terai Y, Ngatunga BP, et al. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. Science. 2015;350: 1493–1498. doi:10.1126/science.aac9927

105. Malinsky M, Svardal H, Tyers AM, Miska EA, Genner MJ, Turner GF, et al. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. Nat Ecol Evol. 2018;2: 1940–1955. doi:10.1038/s41559-018-0717-x

106. Malinsky M, Matschiner M, Svardal H. Dsuite - Fast *D* -statistics and related admixture evidence from VCF files. Molecular Ecology Resources. 2021;21: 584–595. doi:10.1111/1755-0998.13265