Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 2717

Genomics of Neotropical Fruit Flies and Their Maternally Inherited Associates

Konstantinos Papachristos

ACTA UNIVERSITATIS
UPSALIENSIS
2026

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in A1:107a, BMC, Husargatan 3, Uppsala, Friday, 6 November 2026 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English. Faculty examiner: Professor Christoph Bleidorn (Georg-August-University Göttingen).

Abstract

Papachristos, K. 2026. Genomics of Neotropical Fruit Flies and Their Maternally Inherited Associates. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 2717. 93 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2962-8.

This thesis examines the evolutionary history of Neotropical *Drosophila* fruit flies and their maternally inherited associates, i.e., their mitochondria and their intracellular endosymbiont *Wolbachia*. Moreover, we study nuclear genomic features that have emerged from insertions of the mitochondrial genome into the nuclear genome, called Mega-NUMTs.

In the four papers produced for this thesis, we find the following. First, the *Wolbachia* strain *w*Pau that is obligate for its host, i.e., *D. paulistorum*, shows genomic differences in the number of pseudogenes, transposable elements and rearrangements compared to closely related facultative strains. Second, mitochondrial and *Wolbachia* phylogenies are almost perfectly congruent both between and within multiple host species of the willistoni and saltans groups of *Drosophila*, indicating co-speciation since their most recent common ancestor. However, divergence times of mitochondria and *Wolbachia* do not match, suggesting either unprecedentedly slow rates of evolution for a bacterial endosymbiont or horizontal transfer between hosts. Third, we find that *w*Pau from different semispecies of *D. paulistorum* are overall congruent with the phylogeny of their hosts, suggesting co-divergence. These strains also differ in genomic features thought to be involved in host interaction, specifically in a handful of effector proteins, and the genomic arrangement of a genomic island known as the Undecim cluster. Fourth, we find multiple mitochondrial introgression events within the willistoni subgroup, and that one of the two mitotypes of *D. paulistorum*, the β mitotype, is likely introgressed, while the α mitotype may be native to the species. Despite this, we find limited evidence for nuclear compensation in *D. paulistorum*, as a response to the introgression of the β mitotype, indicating that mitochondrial introgression may not have caused deleterious effects. Lastly, we describe tandem insertions of complete "α-like" mitochondrial genomes in chromosome 3 of some lines of *D. paulistorum*, forming Mega-NUMTs. Phylogeny shows they have a common origin, but their genomic location, NUMT completeness and transposable element content supports lineage-specific amplification.

In summary, this thesis contributes to the understanding of the evolutionary relationships between Neotropical *Drosophila*, their mitochondria and their *Wolbachia* endosymbionts.

Keywords: Bioinformatics, Drosophila paulistorum, Evolutionary biology, Genomics, Mega-NUMT, Mitonuclear incompatibility, Neotropical fruit flies, Wolbachia.

Konstantinos Papachristos, Department of Cell and Molecular Biology, Molecular Evolution, Box 596, Uppsala University, SE-752 37 Uppsala, Sweden.

© Konstantinos Papachristos 2026

ISSN 1651-6214
ISBN 978-91-513-2962-8
URN urn:nbn:se:uu:diva-599019 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-599019)

To all my teachers and mentors
that taught me how to think and supported me along the way,
and to all my family and friends that I keep in my heart,
no matter where I am.

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I. **Papachristos, K.**, Miller, W. J., & Klasson, L. (2025). A co-speciation dilemma and a lifestyle transition with genomic consequences in *Wolbachia* of Neotropical Drosophila. *BMC Genomics*, *27*(1), 41.

II. Ekström, E., **Papachristos, K.**, Miller, W. J., Klasson, L. Evolutionary history and genomic diversification of obligate *Wolbachia* in *Drosophila paulistorum*. *Manuscript*.

III. **Papachristos, K.**, Montoliu-Nerin, M., Hua-Van, A., Miller, W. J., & Klasson, L. Lack of co-introgression and compensation after mitochondrial introgression in the willistoni group of Neotropical fruit flies. *Manuscript*.

IV. **Papachristos, K.,** Miller, W. J., & Klasson, L. Shared origin but divergent genomic trajectories of Mega-NUMTs in *Drosophila paulistorum*. *Manuscript*.

Reprints were made with permission from the respective publishers.

Author's contribution to the papers

I. Genome assemblies, genome annotations, comparative analyses, phylogenetic analyses, visualization, original draft.

II. Genome assemblies, genome annotations, methods, contribution to draft.

III. DNA extraction, genome assemblies, annotations, phylogenetic analyses, introgression analyses, visualization, original draft

IV. Annotations, phylogenetic analyses, visualization, original draft

Contents

Artificial intelligence statement 11

Chapter I: *Drosophila* species in the willistoni and saltans groups 13
The *Drosophila* fruit flies 13
Willistoni and Saltans groups 14
Genomes and annotations for the willistoni-saltans groups 18

Chapter II: Symbiosis 19
Symbiosis and endosymbionts 19
Maternally inherited intracellular bacterial endosymbionts 20
Genome evolution of bacterial endosymbionts 23
Wolbachia: A reproductive manipulator 25
Symbiont-induced speciation 28
Wolbachia in Neotropical *Drosophila* 29

Chapter III: Mitonuclear Interactions 31
Mitochondrial functions and mitonuclear interactions 32
Mitonuclear co-evolution 34
Mitochondrial introgression 35

Chapter IV: Nuclear mitochondrial insertions 39
NUMTs and NUMTogenesis 39
From NUMTs to Mega-NUMT 42

Chapter V: Sequencing and Genomics 44
Genome architecture of prokaryotes vs eukaryotes 44
Sequencing technologies 45
An overview of genome assembly and annotation 47
Comparative genomics 50
Some encountered methodological challenges 52

Research Aims 55

Summary of research papers 58
Paper I 58
Paper II 59
Paper III 59
Paper IV 60

Conclusions and perspectives 61
Svensk sammanfattning 63
Ελληνική περίληψη 66
Acknowledgements 69
References 71

Abbreviations

aaRS	Aminoacyl tRNA synthetase
CHCs	Cuticular hydrocarbons
CI	Cytoplasmic incompatibility
DSB	Double strand break
ETC	Electron transport chain
GFP	Green fluorescent protein
ILS	Incomplete lineage sorting
LTR	Long terminal repeats
MRCA	Most recent common ancestor
MSC	Multispecies coalescent
mt	mitochondria-encoded (genes or proteins)
NGS	Next generation sequencing
NHEJ	Non-homologous end joining
N-mt	Nuclear encoded mitochondrial (genes or proteins)
NUMT	Nuclear mitochondrial transfer
OXPHOS	Oxidative phosphorylation
qPCR	Quantitative polymerase chain reaction
SINE	Short interspersed nuclear element
SNP	Single nucleotide polymorphism
TE	Transposable element

Artificial intelligence statement

Various versions of ChatGPT, from 4o to 5.6 Sol, have been used to create code used in the papers of this work. All scripts have been manually checked to verify the correctness of their outputs. Moreover, ChatGPT has been used as a search tool for scientific publications, some of which have been used are references for this thesis. All such references have been read by the author to verify their adherence to the claims made in this thesis. ChatGPT was used to produce part of Figure 3, which was later modified to fit the needs of the author.

Chapter I: *Drosophila* species in the willistoni and saltans groups

This chapter provides a general overview of the systematics and biology of fruit fly species belonging to the genus *Drosophila*. It then focuses on the two understudied groups of Neotropical *Drosophila* species that are the subject of this thesis, the willistoni and saltans groups. Finally, I review the genomic information currently available for these two groups.

The *Drosophila* fruit flies

Drosophila is a genus of saprophagous flies and comprises more than 1,500 species in nine subgenera [1]. The largest subgenera are Drosophila and Sophophora, together accounting for more than 90% of the species in the *Drosophila* genus [2]. Species in the subgenera are further divided into species groups and subgroups. The phylogenetic relationships of the groups have recently been comprehensively identified with the publication of hundreds of genomes from species representatives from all groups of *Drosophila* and show that the genus is paraphyletic [3, 4]. Molecular dating based on three fossils of diptera, one fossil of Eocene Drosophilidae and the age of the Hawaiian Islands, place the age of the *Drosophila* genus at around 50-56.6 million years and the Sophophora-Drosophila subgenera divergence at around 43.9-49.9 million years [5]. Moreover, the evolutionary history of the genus is complex, as at least 30 pairs of species within the *Drosophila* genus have hybridised, with some pairs sharing more than 10% of their genomes [5].

Drosophila biology is a generally well understood topic. Flies of the *Drosophila* genus feed on yeast and bacteria of decomposing plant matter, such as that of fruits, flowers, leaves, bark, stem and roots, on slime fluxes of trees and on flesh fungi [6]. Yeast in the diet of *Drosophila* is of great importance as they are the main source of protein in the flies' fruit-based diets, which are limited in protein [7, 8]. Moreover, specific yeast species affect survival, development time and adult body weight [9]. Bacteria are also important modulators of feeding behaviour in *Drosophila*, as commensal gut bacteria in *D.*

melanogaster can modulate feeding behaviour by increasing appetite for yeast and increase egg laying upon deprivation of essential amino acids [10].

The life cycle of *Drosophila* species includes four developmental stages, i.e., egg, larva, pupa and adult, and is characterized by complete metamorphosis [6]. For most *Drosophila* species, adult flies are attracted to their feeding and breeding sites by the volatile compounds of fermenting matter, though for some species feeding and breeding sites are not the same. After mating, females proceed to look for an appropriate oviposition site and lay their eggs either on or inside the substrate, depending on the species. The oviposition sites are chosen by the females according to environmental cues such as ambient light, plant host chemistry and texture, microbial composition, temperature and the presence or absence of larvae. Larvae feed on the decaying matter and then undergo pupation either in their larval food, or the nearby ground. From pupae emerge sexually immature adult flies, which will mature in a span of days to several weeks, depending on the species. Sexually mature adults then employ olfaction, auditory, tactile and visual signals and proceed to courtship behaviour [11].

Willistoni and Saltans groups

This thesis examines species from two species groups within the subgenus Sophophora, the willistoni and saltans groups, with a primary focus on the willistoni group. The willistoni and saltans groups form a monophyletic Neotropical clade [12] that is sister to species from the *Lordiphosa* genus [1, 4], thus making Sophophora a paraphyletic group. The Neotropical Sophophora-Lordiphosa clade is sister to the old-world melanogaster and obscura groups of *Drosophila*.

The saltans group includes 22 species in five subgroups, the sturtevanti, the saltans, the parasaltans, the elliptica and the cordata groups [13]. Most species have a rather narrow distribution in the Neotropics, with notable exceptions being *D. prosaltans* and *D. sturtevanti*, which are distributed across central and south America (Figure 1).

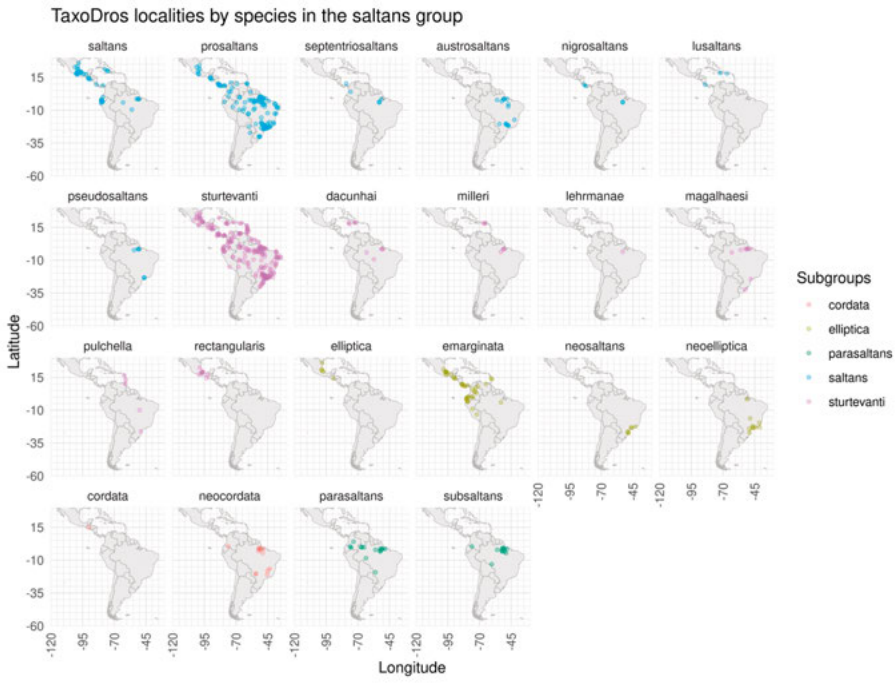

Figure 1. Species distribution for the 22 species of fruit flies in the saltans group according to collection sites in TaxoDros [13]. Collection sites for each species are coloured according to their subgroup.

The willistoni group includes 24 species in three subgroups, the willistoni, the bocainensis and the alagitans subgroups [13–15]. Some species are understudied endemic species with narrow distribution, while others have a more widespread distribution in central and south America and the Caribbean islands (Figure 2). Specifically, the willistoni subgroup includes six species, *D. paulistorum*, *D. pavlovskiana*, *D. equinoxialis*, *D. willistoni*, *D. tropicalis* and *D. insularis*. Two species, *D. pavlovskiana* [16] and *D. insularis* [17], are narrow endemics with the former being in Guyana and the latter in some islands of the Lesser Antilles. The remaining species have a broader sympatric distribution in central and south America, with *D. willistoni* having the broadest distribution in the willistoni subgroup (Figure 2).

A recent study on the broader phylogeny of the *Drosophila* genus, estimated the age of the willistoni-saltans clade to be between 14.2 and 20.6 million years [5]. The relationship between the species of the willistoni and saltans groups can be seen in figure 3. In the willistoni subgroup, the first species to diverge from the rest is *D. insularis*, then follows *D. tropicalis*, followed by divergence of *D. willistoni* and finally, *D. equinoxialis* and *D. paulistorum* are sister species. Molecular data on *D. pavlovskiana* do not exist to confidently

place the species on the tree, however, given that it was originally described as a "Guianan form" of *D. paulistorum* [16], it could be that it is a close relative *D. paulistorum*.

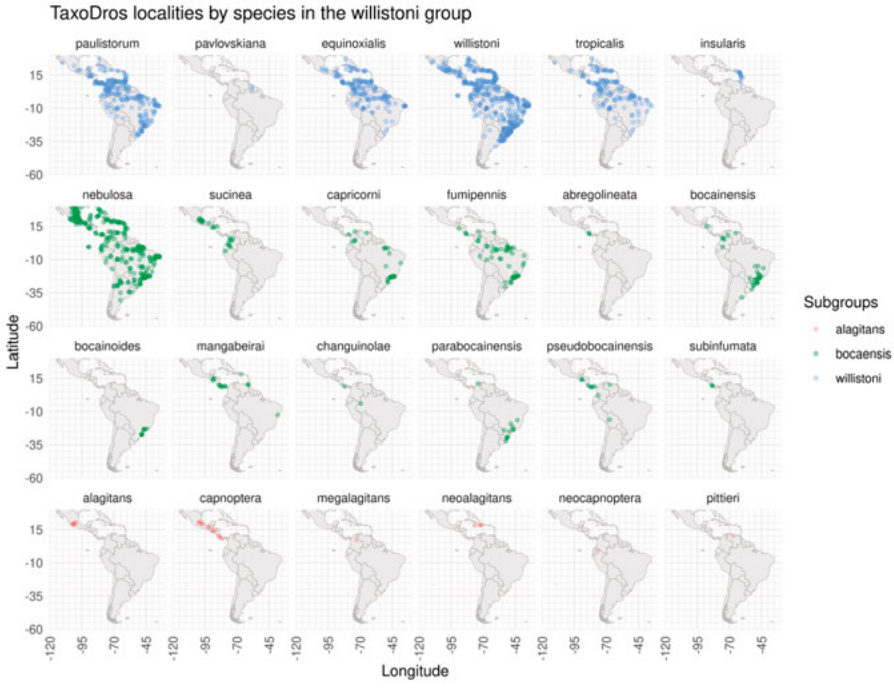

Figure 2. Species distribution for the 24 species of fruit flies in the willistoni group according to collection sites in TaxoDros [13]. Collection sites for each species are coloured according to their subgroup.

Most species in the willistoni subgroup include subspecies which exhibit partial unidirectional or bidirectional incompatibility with each other, as males produced are sterile [18–20]. More interesting is the case of *D. paulistorum* undergoing speciation, where six partially incompatible semispecies are recognised, the Orinocan (OR), the Amazonian (AM), the Andean-Brazilian (AB), the Centroamerican (CA), the Interior (IN) and the Transitional (TR) semispecies [21, 22]. These semispecies are both prezygotically isolated, as they exhibit assortative mating [23–25] and postzygotically isolated, as few inter-semispecies crosses produce progeny that are sterile [22, 26]. Assortative mating in the semispecies is related to differentially produced cuticular hydrocarbons (CHCs) in ratios and quantities [27]. Interestingly, both prezygotic and postzygotic isolation seem to be partially driven by the endosymbiont of *D. paulistorum, Wolbachia*. CHC profiles are altered in *Wolbachia* knockdown males of *D.*

paulistorum and those individuals are rejected by wild type females [28], while fecundity is reduced in female *Wolbachia* knockdowns [29].

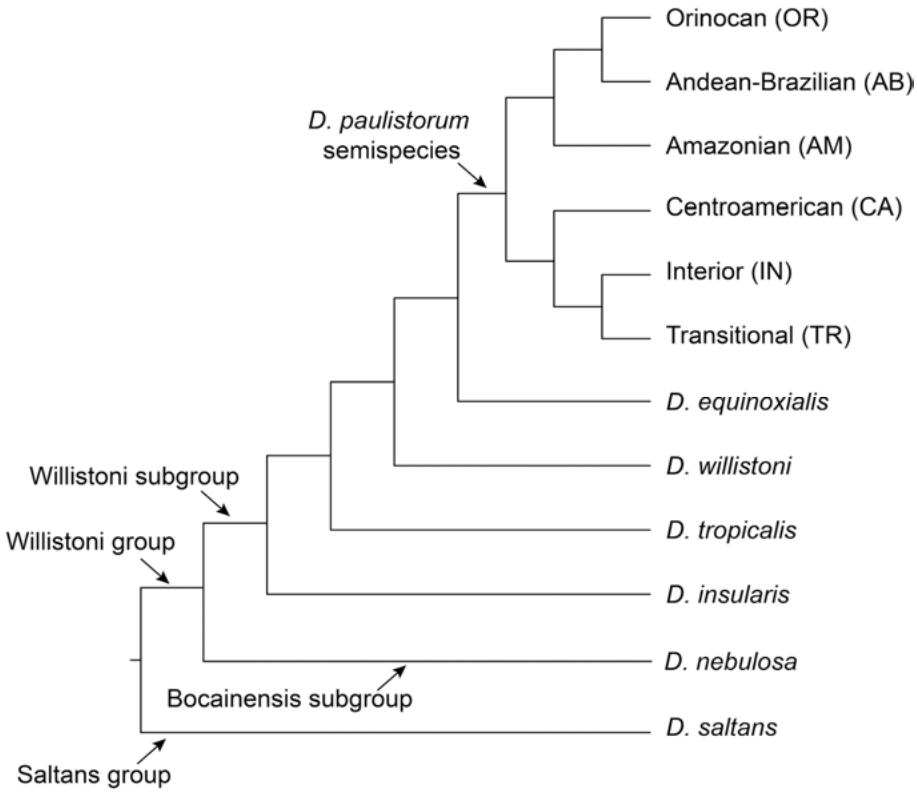

Figure 3. Cladogram showing the phylogenetic relationships of species in willistoni subgroup [5, 30]. *D. nebulosa* and *D. saltans* are shown to as representatives of the bocainensis subgroup and saltans group respectively. Positions on the cladogram for *D. pavlovskiana* and species from the alagitans subgroup are uncertain.

Knowledge on the ecology and life history of the willistoni and saltans group species is restricted. Species of the willistoni group have been reported in 114 plant species and the saltans group in 25 [31]. However, many of the fruit flies could not be identified to the species level, while the microbial content of the plant hosts reported is not known. It is unclear whether species with a high number of plant hosts, such as *D. willistoni* and *D. paulistorum*, really are generalists, or whether all those plant hosts share specific microbes that are attractive to the flies. Life history traits have been studied in a handful of species, namely *D. willistoni* [32] and *D. saltans* and *D. austrosaltans* [33]. Finally, a few studies look at phenotypic plasticity of *D. willistoni* [34] and the effects of human disturbance on its abundance [35]. To conclude, most species of the willistoni and saltans groups are understudied in nature.

Genomes and annotations for the willistoni-saltans groups

Modern sequencing techniques have allowed researchers to gather high-quality genomic data and produce high-quality genome assemblies for the willistoni and saltans groups [4] (Paper III). Additionally, with the Neotropical fruit flies being closely related to *D. melanogaster*, good quality gene annotations for their genomes can be produced. Annotations require mRNA data from the same species from which the assembly was made to provide as gene evidence during the annotation process. Until recently, these types of data were not available for most Neotropical species, but this trend is shifting, with more genome assemblies accompanied with gene annotations being published.

The first Neotropical *Drosophila* genome came from *D. willistoni* and was published as part of the "Drosophila 12 genome consortium" [36]. To my knowledge, this was the only Neotropical *Drosophila* genome available until 2021, and since then more genomes are added in public databases every year. Kim and co-authors [4] assembled 101 genomes with long Nanopore reads from 14 species groups of *Drosophila*, including species from the willistoni and saltans groups, but annotations were not available. Sottolano and co-authors [37] provided an annotated genome for *D. nebulosa* of the willistoni group, bocainensis subgroup. Baião and co-authors [38] assembled draft genomes from several species from the willistoni and saltans groups, as well as complete mitochondrial genomes, but the draft nuclear assemblies are fragmented and not annotated. Prediger and co-authors [39] assembled draft genomes from 15 species of the five saltans subgroups and several mitochondrial genomes, but the draft nuclear genomes are fragmented and not annotated. Gebert and co-authors [40] produced mRNA and Hi-C data to create highly contiguous annotated genomes for several *Drosophila* clades using data from [4], including species from the willistoni group. Finally, Dhakad and co-authors [41] annotated RefSeq reference genomes using mRNA data for 301 *Drosophila* species, including species from the saltans group.

Thus, at the time of writing this thesis, two contiguous genomes for *D. paulistorum* are available, one of which belongs to the AB semispecies (the L12 line), and the other (L06) is of uncertain origin [4]. Also, contiguous genomes for *D. equinoxialis*, *D. willistoni*, *D. tropicalis*, *D. insularis*, *D. sucinea*, *D. nebulosa*, *D. saltans*, *D. prosaltans*, *D. austrosaltans*, *D. neocordata* and *D. sturtevanti* are available. Most of the saltans group species have fragmented draft genomes and most of the bocainensis and alagitans subgroup species have no available genomic data.

Chapter II: Symbiosis

This chapter introduces the concepts of microbial symbiosis and endosymbiosis, with particular emphasis on maternally inherited intracellular bacterial endosymbionts. It then discusses the evolutionary patterns expected to arise between these endosymbionts and their hosts, as well as endosymbiont genome evolution and how it is shaped by endosymbiont lifestyle and host dependence. Next, I provide a general introduction to *Wolbachia*, the endosymbiont of interest for this thesis. Finally, I briefly review what is known about *Wolbachia* infections in Neotropical fruit flies of the willistoni and saltans species groups.

Symbiosis and endosymbionts

Symbiosis is the "*phenomenon in which dissimilar organisms live together*", as defined by Heinrich Anton de Bary in 1879 [42]. Symbiotic relationships are important for the health of ecosystems, as they facilitate vital ecological functions, and are an interesting subject for evolutionary studies, as organisms must reciprocally adapt to their symbiotic partners [43]. Additionally, symbiotic relationships are diverse, as they include multiple partners from all domains of life, prokaryotes, archaea, eukaryotes and viruses. Particularly, bacterial symbioses, the symbioses between bacterial symbionts and their eukaryotic host, are numerous in insects and other animals [44].

Describing bacterial symbioses is not a trivial task, as symbionts form complicated and diverse relationships with their hosts. Bacterial symbioses can be classified in multiple ways, with one being the nature of the host-symbiont relationship [45]. A symbiosis will be *facultative* when either the relationship is not necessary for the survival of the host or the symbiont. *Obligate* symbiosis means the relationship is mandatory for either host or symbiont survival. Whether these definitions refer to the host or the symbiont depends on context and the field [46]. For example, in the field of bacterial symbiosis usually the facultative or obligate refer to the relationship in respect to the host, i.e., if the host can survive without its symbiont or not. In symbiosis of protists, facultative or obligate refer to the symbiont, i.e., whether the protist can survive

without a host or not. In this work, the terms facultative or obligate will refer to the relationship of the endosymbiont to its host, and any alternative use will be explicitly indicated.

Another way to characterise symbioses is based on the location of the symbiont [47]. *Ectosymbiosis*, is when the symbionts live outside or on the surface of their hosts. In *endosymbiosis*, the symbionts live inside the body of their hosts. *Intracellular* endosymbionts, live inside the cells of its host and *extracellular* endosymbionts live within host tissues but not inside its cells. A *mixed* category of endosymbionts also exists, in which they have both extracellular and intracellular stages in their life cycle [45].

Categorisation can also be based on the transmission of symbionts. *Horizontally* transmitted symbionts are picked up by their hosts from the environment or from other host species, while *vertically* transmitted symbionts are inherited from one generation of hosts to the next, usually by the mother [48].

Finally, based on how the relationship affects the fitness of the host, symbiosis can be *parasitic*, where fitness of the symbiont increases at the expense of its host, *commensal* where host fitness is not affected by the symbiont, and *mutualistic*, where host fitness increases [45]. However, this definition of mutualism may not satisfy everyone, as it does not explicitly require the symbiont to have an adaptive effect on its host. There are cases where bacterial symbionts facilitate "endogenous" functions of their hosts, such as feeding habits or ovarian development, and these may be viewed as cases of evolutionary addiction, rather than cases of mutualism [49]. In those, removal of the symbiont will cause a reduction in the fitness of the host and therefore will technically satisfy the described definition of mutualism. However, the host's fitness is lower than what one would expect, reflecting a loss of function in the host, rather than an adaptive benefit by the symbiont. For this work, I will use mutualism as defined above, without necessarily implying adaptive effects on the hosts.

Maternally inherited intracellular bacterial endosymbionts

One of the research areas in the field of bacterial symbiosis is the study of intracellular bacterial endosymbionts. These are usually transmitted vertically via the mother but may occasionally transfer horizontally into new host species. The way vertical transmission occurs is specific to the symbiont, as some transmit via the egg, while others transmit via alternative routes, such as coating the embryo with symbiont cells or complex transfer pathways involving the midgut followed by exocytosis and endocytosis [50]. Some intracellular endosymbionts may have some non-intracellular life stages and

can thus live outside host cells and be grown in cell lines and/or axenic media, i.e., media that contain only the endosymbiont of interest, without contamination by any other living organism. There are other endosymbionts that cannot live outside of host cells, and thus can only be cultured in cells lines or inside of their hosts [45].

In the context of mutualistic relationships with maternally inherited intracellular bacteria, the endosymbionts are described by further terminology that captures the configuration of endosymbiosis in a host [44, 46]. *Primary* endosymbionts are mutualistic and obligate for the survival of the host. These endosymbionts are expected to have been long-term associated with their hosts, potentially for tens to hundreds of millions of years [51] and to follow their host's phylogeny [46, 52, 53], i.e., endosymbionts co-speciate with their hosts (Figure 4). It is usual that in these cases the endosymbionts provide essential nutrients to their hosts [54]. Thus, for many insect hosts, these symbiotic relationships are so deep, and vital for their survival that they have evolved specialised body structures to host their primary endosymbionts. These host structures can be specialised host cells called bacteriocytes in which the endosymbionts reside, or specialised organ-like structures called bacteriomes which contain multiple bacteriocytes. Bacteriocytes and bacteriomes have evolved multiple times in hosts of many insect orders. Some examples include Hemiptera where Aphids have *Buchnera* in bacteriomes [55], in Blattodea which have *Blattabacterium* in bacteriocytes/fat-body-associated tissues [56], Coleoptera, where some weevils and other beetles have bacteriocyte-associated nutritional symbionts [57], in Diptera where Tsetse flies have bacteriome-associated *Wigglesworthia* [58], and Phthiraptera where some lice have bacteriomes or bacteriocyte-like symbiotic organs [59].

Secondary endosymbionts are all endosymbionts that are inherited along with the primary symbiont but are not needed for growth and survival of their hosts, i.e., they are facultative for their host [44, 60]. They may be mutualistic, providing to the host benefits that depend on their environment, or they may be parasitic. In some cases, secondary endosymbionts complement the loss of function in primary symbionts, thus rescuing the symbiosis and becoming *co-primary* endosymbionts, i.e., mutualistic and obligate for their hosts [61, 62]. Co-primary and secondary endosymbionts are usually short-term associated with their hosts and usually do not follow the pattern of the hosts' phylogeny (Figure 4) as they have been horizontally transferred [46, 60]. However, some co-primary endosymbionts have been shown to have co-diverged with their hosts [62], indicating that rescue of the symbiosis can occur in host ancestors before their diversification. Additionally, co-primary and secondary symbionts are prone to replacement following changes in host diets, host ecological

niche, or function of the host's genome [61, 62]. Secondary symbionts may invade and compete with obligate symbionts for access to the bacteriocytes of hosts [63], sometimes even completely replacing a co-primary symbiont [53, 62, 64]. In *Bemisia tabaci* whiteflies, bacteriomes are shared with their primary symbiont *Portiera*, and several other symbionts belonging to *Hamiltonella*, *Arsenophonus*, *Cardinium*, *Wolbachia* and *Rickettsia* [63]. In aphids of the subfamily Lachninae, the co-primary *Serratia symbiotica* strain has been replaced by endosymbionts belonging to *Sodalis*, *Erwinia* and unnamed Enterobacteriaceae strains [64]. In cicadas of the *Chremistica ochracea* species that harbour *Hodgkinia* and *Karelsulcia* as their primary endosymbionts, a yeast-like symbiont (YLS) seems to be replacing *Hodgkinia*, and *Hodgkinia* co-localises with *Karelsulcia*, presumably to escape replacement by YLS [53]. In phyllids of *Cacopsylla pyrisuga*, *Sodalis* strains have been acquired multiple independent times and have replaced the co-primary symbiont *Psyllophila* [62]. Thus, replacement of co-primary endosymbionts, rather than being adaptive, may be driven by population dynamics of the competing facultative and co-primary endosymbionts [53, 64].

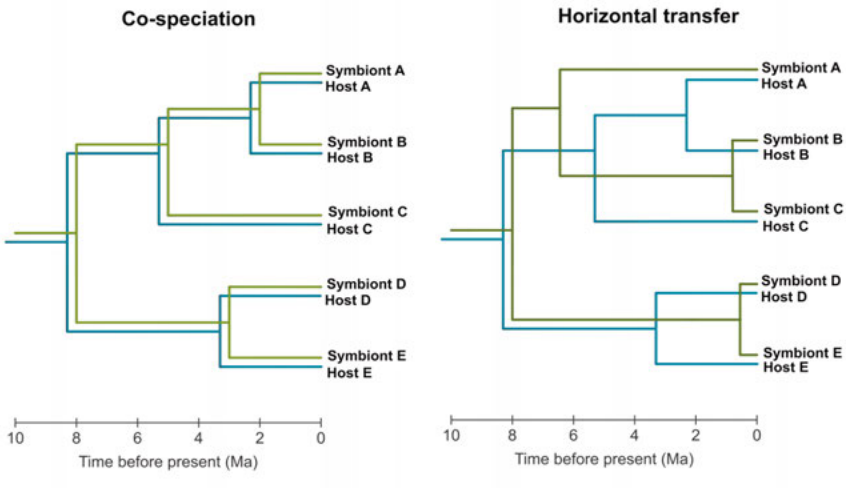

Figure 4. Phylogenetic patterns between hosts and their endosymbionts. Primary endosymbionts, on which their hosts depend for survival, have co-diverged with their hosts over long periods of time. This is because speciation events of the endosymbionts are a result of speciation events of their hosts, and thus the two partners have been co-speciating. Co-primary and secondary endosymbionts that have been recently acquired by their hosts via horizontal transfer, will not have their phylogenies mirror that of their hosts'. Both topology and divergence times are expected to mismatch in this case. Host's and endosymbionts' phylogenies are in blue and green respectively. Part of this image was created using ChatGPT Sol 5.6 and was then edited in Inkscape.

Some maternally inherited intracellular bacteria are parasitic by affecting the reproduction of their hosts. Endosymbionts such as *Wolbachia*, *Spiroplasma*, *Arsenophonus* and *Cardinium* [65], induce various reproductive manipulation phenotypes, which promote the spread in their hosts' populations. These phenotypes may be Cytoplasmic Incompatibility, Male-killing, Feminization and Parthenogenesis [66]. Horizontal transfer of reproductive manipulators between host species is quite common [67–69], but the mechanisms involved are not understood, though they may include transfer between predator-prey, between parasitoid-host, or even plant to insect [70].

It is important to note that bacterial endosymbioses are dynamic and may change in nature during evolutionary time. Symbionts can transition towards any direction of the symbiosis spectrum during their evolutionary history, from parasites to mutualists or vice versa, though the latter is uncommon [71]. This may have to do with the fact that hosts usually exert control over their bacterial endosymbionts [72, 73]. Density of endosymbionts is controlled by the host, depending on the host's nutritional needs and development [74, 75], first by confinement of the symbiont in specific areas of the body, like bacteriocytes and bacteriomes. Second by immune response [76] and autophagy [73, 77], or other types of cell lysis [78], in cases where endosymbionts invade other host tissues [76]. Thus, it may be rare for long-term mutualists to escape host control and transition to parasitism [79]. A third reason might be that the genomes of long-term endosymbionts usually undergo extreme reduction [80], as will be explained later on this chapter, which would make the endosymbiont completely dependent on its host for performing vital functions. On the other hand, it is more common for endosymbionts that are considered reproductive parasites to offer some adaptive advantage to new hosts they encounter. For example, *Spiroplasma* can protect *D. melanogaster* from bacterial and yeast pathogens by boosting nutritional immunity and melanization [81] and can also recover fertility in *D. neotestacea* after infection from its parasitic nematode *Howardula aoronymphium* [82]. In another example, the *Wolbachia* strain *w*MelCS strain increases fecundity in *D. melanogaster*, compared to uninfected flies [83] and *Wolbachia* is well known to provide viral protection to its hosts [84, 85].

Genome evolution of bacterial endosymbionts

The genome of vertically transmitted intracellular endosymbionts follows a reductive trajectory that is related to the lifestyle of the endosymbiont. As bacterial endosymbionts evolve from free-living to host-restricted bacteria, they experience genome degradation due to small effective population sizes and

the inherent bias of bacterial genomes for deletions [80]. Moreover, only a portion of the endosymbionts are transmitted from mother to offspring in each generation, creating bottlenecks [86] that further reduce the effective population size of endosymbionts. Due to the low effective population size, populations are more exposed to random fluctuations in allele frequencies from genetic drift and thus natural selection becomes inefficient in removing deleterious mutations or deletions [87]. Finally, host restriction limits gene exchange and recombination with other endosymbionts, reducing opportunities to facilitate the removal of deleterious mutations [88, 89]. This eventually leads to the accumulation of deleterious mutations in bacterial chromosomes through a process known as *Muller's ratchet* [90]. Thus, genomes of maternally inherited endosymbionts tend to accumulate deleterious mutations and reduce in size [80].

Genome reduction in endosymbionts depends also on the degree with which its host depend on it for survival, i.e., *host dependence*. In bacterial symbionts from hosts across animals, plants and fungi, genome size of vertically transmitted symbionts decreases as host fitness, a proxy for host dependence, increases [91]. However, this correlation is not significant in horizontally transmitted symbionts. Moreover, host dependence is higher when the association is obligate rather than facultative for the symbiont, suggesting that symbiont and host dependence are linked [91].

Stages of genome degradation correlate with the time of host-restriction [80]. Host-restriction refers to the degree which a symbiont needs a host to survive. Recently host-restricted endosymbionts are facultative, have smaller genomes than their free-living relatives due to deletions, have accumulated more transposable elements and pseudogenes, and have more rearranged chromosomes. These happen because host-restriction first allows for the symbiont to depend on its host for some of its functions that are thus no longer maintained by negative selection, and second it reduces the effective population size of the endosymbionts, making selection ineffective at keeping functions that are non-vital. Long-term host-restricted endosymbionts are obligate and compared to their recently host-restricted relatives, they have smaller genomes, fewer transposable elements and pseudogenes, and chromosomal rearrangements are rare, as genome reduction progresses further with time [80]. For example, obligate endosymbionts of the genus *Sodalis* have the smallest genome sizes, followed by their facultative and then by their free-living relatives [92].

Given that some endosymbioses are estimated to have lasted up to hundreds of millions of years [51], it would be interesting to know how endosymbionts can escape Muller's ratchet until genome degradation becomes too great and

causes extinction of the endosymbiont. Some proposed ways include recombination and gene-conversion [93], horizontal gene transfer [94], and recruitment of other co-primary symbionts [53, 64]. Another possibility is an increase in genome copy number per endosymbiont cell [95]. It is not known if endosymbionts have high genome copy numbers universally, but in a few cases where the topic has been studied, it is suggested that it can be high. In *Buchnera aphidicola*, genome copy number per endosymbiont cell was estimated to range from 10 to 200 copies [96]. In *Carsonella*, it is estimated that a single cell can have thousands to tens of thousands of copies of its genome [95]. A high copy number per endosymbiont cell could allow for removal of deleterious variants by favouring non-mutant copies as templates for DNA repair and thus reduce the effective substitution rate in the endosymbiont's genome [95].

Wolbachia: A reproductive manipulator

The bacterial endosymbiont of interest for this thesis is *Wolbachia*. *Wolbachia* is an alpha-proteobacterium that is an endosymbiont of arthropods and nematodes [97, 98]. It was first described in the mosquito *Culex pipiens* [99] as one species, *Wolbachia pipientis,* but several phylogenetically distinct clades called "supergroups" have been identified, which could potentially be considered distinct *Wolbachia* species [100]. *Wolbachia* is vertically transmitted via the egg from mother to progeny and is an obligate intracellular endosymbiont [101]. Thus, *Wolbachia* has not been cultured in axenic media [45]. This means that genetic engineering tools, such as plasmid transformation, are not available for studying *Wolbachia* functions. Experimental approaches available for studying *Wolbachia* are cell cultures from either naturally or artificially infected *Wolbachia* hosts, such as cell lines from diptera, lepidoptera and ticks [102], transinfection experiments in host lines [103], and use of antibiotics to study the effect of *Wolbachia* on its hosts [29].

Wolbachia is known to be mainly a reproductive parasite. Its spread is attributed to the reproductive manipulation phenotypes it induces in its hosts, namely, cytoplasmic incompatibility (CI), male-killing, parthenogenesis and feminization. As *Wolbachia* is maternally transmitted, these phenotypes increase the fitness of individuals able to transmit the endosymbiont, i.e., infected females, by reducing the fitness of non-transmitting individuals, i.e., uninfected females or males [104]. Interestingly, a prophage insertion in *Wolbachia* genomes, the WO prophage, seems to be important for reproductive manipulation, as it is found in many *Wolbachia* genomes, except obligate mutualistic strains [105–107]. Also, the WO prophage carries an area called

Eukaryotic Association Module (EAM) that contains genes with eukaryotic features, implying transfer between prokaryotes and eukaryotes via the prophage [106]. In EAM, proteins have been identified to induce all four reproductive manipulation phenotypes [108–112], or control *Wolbachia* titer and virulence [107, 113, 114]. Notably, a set of 11 genes related to virulence, called Undecim cluster, is reported to occur only in one copy per *Wolbachia* genome in EAM of some WO prophages, and may be involved in *Wolbachia*-host interactions [107].

The most common of the reproductive phenotypes is CI which induces embryonic lethality in crosses of uninfected female hosts and infected male hosts [100]. CI can be distinguished between unidirectional and bidirectional (Figure 5). Unidirectional CI occurs when incompatibility is expressed in only one direction of a cross, either when *Wolbachia*-infected males mate with uninfected females, or when hosts carry different *Wolbachia* strains and one strain can rescue the CI induced by the other, but not vice versa. In contrast, bidirectional CI occurs when hosts carry two different *Wolbachia* strains that are mutually incompatible, such that neither strain can rescue the CI induced by the other [100].

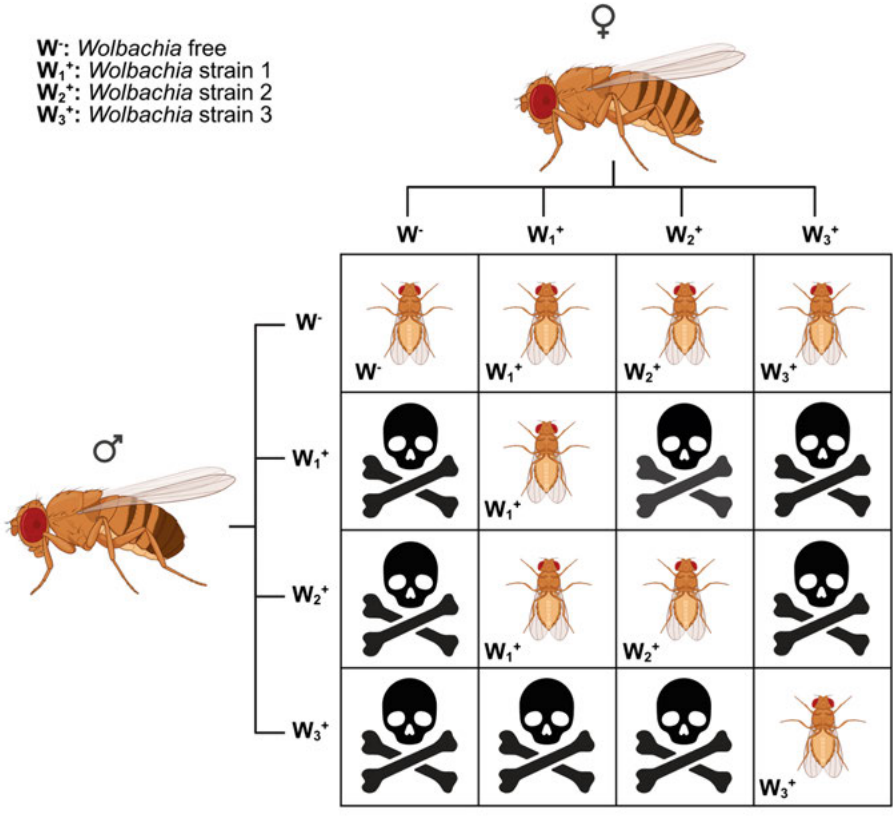

Figure 5. Cytoplasmic incompatibility caused by *Wolbachia*. *Wolbachia*-infected males that mate with *Wolbachia*-free females produce no progeny due to CI, while *Wolbachia*-free males that mate with *Wolbachia*-infected females produce progeny. Thus *Wolbachia*-infected males are unidirectionally incompatible. When both parents are infected by different *Wolbachia* strains, CI may be rescued in one cross but not the reverse, as is in the case of females infected by W_1 strain rescuing incompatibility in W_2 males. In this case, W_2 is unidirectionally incompatible to W_1. If no *Wolbachia* strain can rescue CI caused by the other, as is in the case of W_3 when crossed in any direction with either W_1 or W_2, then CI is bidirectional. *Created in BioRender. Papachristos, K. (2026) https://BioRender.com/ozcvoh6*.

The underlying mechanism for the embryonic lethality in CI is described in two candidate models, the host modification model where *Wolbachia* modifies the sperm [115] and the toxin-antidote model where *Wolbachia* secretes a toxin in the male sperm [116]. Depending on the model, the modification is either reversed, or the antidote is provided in *Wolbachia* infected eggs, thus rescuing the cross of *Wolbachia* infected males with *Wolbachia* infected females. Two proteins have been identified to be involved in CI [108, 109], CifA which participates in both embryonic lethality and its rescue, and CifB which participates in embryonic lethality.

Conventionally, *Wolbachia* presence is detected by PCR targeting fragments from the *wsp Wolbachia* gene [117], or by using a multilocus sequencing typing (MLST) PCR scheme from fragments of five other *Wolbachia* genes [118]. However, sometimes *Wolbachia* infections are at such low titers in the body of the host that it may reach densities below the detection threshold of the standard MLST or *wsp* PCR methods, requiring new targets [119]. Despite these difficulties in detection, *Wolbachia* is estimated to infect up to 60% of all insect species [120]. Given this global epidemic of *Wolbachia*, it is not surprising that horizontal transfer between distantly related hosts is quite common, and thus, when looking at host-*Wolbachia* phylogenies at a large scale, co-divergence of insects and their *Wolbachia* is not observed [69]. However, *Wolbachia* has been shown to co-diverge with their hosts in cases where it is an obligate mutualist, such as in nematodes infected by *Wolbachia* of the C and J supergroups [121], and in bedbugs [122], but also in a few cases where *Wolbachia* is facultative in wasps [123] and solitary bees [124]. The timeframe in which it is usual for *Wolbachia* to co-diverge with its hosts is not known. As no direct measurements of the *Wolbachia* mutation rate are available to this day, it is difficult to date *Wolbachia* infection events. Therefore, when comparing phylogenies of a handful of closely related hosts and closely related *Wolbachia* strains, it is difficult to discern cases of co-divergence since host speciation from cases of independent *Wolbachia* acquisition via horizontal transmission. Some indirect estimates for how long *Wolbachia* infections last in species come from the number of mitochondrial substitutions that have

accumulated in the mitochondrial gene COI [125]. If these substitutions are used as a proxy for time, it is estimated that uninfected arthropod lineages acquire *Wolbachia* every 9.3 million years, and infected lineages lose *Wolbachia* every 7 million years. Thus, it may be possible for *Wolbachia* to co-diverge with its hosts, if the latter speciated a few million years ago.

Symbiont-induced speciation

Speciation is thought to occur with the establishment of several barriers to gene flow, with some barriers promoting the establishment of more [126]. CI-inducing and sex distorting endosymbionts can theoretically restrict gene flow between populations of their hosts under special circumstances [127, 128] and thus could act as initial barriers of the speciation process of their hosts or promote the isolating effect of already existing barriers [129]. Empirical observations also support this. For example, reproductive isolation of two species of parasitoid wasps, *Encarsia suzannae* and *E. genarroi*, is attributed partially to intrinsic incompatibilities and partially to CI caused by *Cardinium* infection, with the former causing up to 20% hybrid lethality and the latter increasing hybrid lethality in the same crossings by 60% [130]. In two subspecies of the butterfly *Danaus chrysippus*, male-killing *Spiroplasma* drives their separation by creating a "genetic sink" in their hybrid zone, where *Spiroplasma*-free males of the dorippus subspecies migrate to mate with S*piroplasma*-infected females of the chrysippus subspecies, resulting in the killing of hybrid males while hybrid females end up with a neo-W chromosome that is presumably linked to colour pattern and *Spiroplasma* susceptibility [131].

The aforementioned examples of endosymbionts play a part in postzygotic isolation of their hosts, but endosymbionts may also participate in prezygotic isolation. In the hybrid zone of *D. recens*, which is infected by a CI-inducing *Wolbachia*, and the closely related *D. subquinaria*, which is not *Wolbachia* infected, female *D. subquinaria* mate with *D. recens* less frequently in sympatry than in allopatry, while female *D. recens* do not show much difference in mating with *D. subquinaria* in sympatry or allopatry [132]. This is due to the cost of interspecies unidirectional CI, where *D. subquinaria* females mate with *Wolbachia*-infected *D. recens* males while in sympatry. In another example, *Wolbachia* affects the length of the spermathecal duct in the *Allonemobiussocius* complex of crickets, a trait thought to be relating to reproductive isolation in insects [133], and as mentioned in Chapter I, *Wolbachia* affects the pheromone profiles of its *D. paulistorum* hosts affecting mate choice [28]. Thus, endosymbionts may contribute to the initial stages of speciation of their hosts, by introducing various pre or postzygotic barriers to gene flow.

Wolbachia in Neotropical *Drosophila*

The Neotropical *Drosophila* of the willistoni and saltans groups are infected by closely related *Wolbachia* strains that are related to the *w*Au strain from *D. simulans* [134, 135]. *D. simulans* is a fruit fly with a worldwide distribution belonging to the melanogaster group and seems to have acquired the *w*Au strain via horizontal transfer from *D. tropicalis* [135]. *Wolbachia* strains that are closely related to *w*Au, have been identified to infect species of the willistoni subgroup, *D. willistoni* (*w*Wil strains), *D. tropicalis* (*w*Tro strains), and the saltans group, *D. prosaltans* (*w*Pro), *D. septentriosaltans* (*w*Spt) via *wsp* PCR [134, 135]. *Wolbachia* has also been shown to infect *D. equinoxialis* (*w*Equ strains) and *D. insularis* (*w*Ins strains) via qPCR, but it is unknown if these strains are related to the other Neotropical *Wolbachia* strains, as *w*Equ and *w*Ins infections are in low-titer [135] and DNA produced for sequencing is not enough. *D. paulistorum* has also been shown to be infected by *Wolbachia* (*w*Pau strains) [29], previously identified as mycoplasma-like organisms [136] that are found in all individuals of *D. paulistorum* [137] and that induce male hybrid sterility between the semispecies [138, 139]. Also, *Wolbachia* affects the CHC profile of *D. paulistorum*, affecting mating choice [28]. Moreover, diversity in infection titer of *w*Pau strains has been shown, with some lines from the OR semispecies having high-titer infections, while other lines from the AM and CA semispecies having low-titer [119]. Because low-titer infections produce little *Wolbachia* DNA for sequencing, so far, it has been difficult to identify the nature of the low-titer *Wolbachia* infections in the Neotropical fruit flies.

Several Neotropical strains have lost both or one of the genes responsible for CI. Specifically, the *w*Wil and *w*Tro strains of *Wolbachia* from *D. willistoni* and *D. tropicalis* respectively, have the *cif* genes absent from their genomes [140]. Other Neotropical *Wolbachia* strains that infect fruit flies outside the willistoni-saltans group, *w*Bor, *w*Inn and *w*Inc, also carry frameshift mutations in their *cifB* genes (Paper I), suggesting that most strains of the Neotropical clade do not cause CI. The only exception could be *w*Pro infecting *D. prosaltans* and *w*Pau, both of which have the *cif* genes (Paper I). Positive fitness effects on its hosts have only been demonstrated in the obligate *w*Pau from *D. paulistorum* [29], and thus the remaining strains closely related to *w*Au in the willistoni and saltans groups, are assumed to be facultative endosymbionts.

Genome assemblies from *Wolbachia* strains of the willistoni and saltans groups have only recently become available. These include draft assemblies of *w*Tro (strain of *D. tropicalis*) and *w*Stu (strain of *D. sturtevanti*) [140], a complete genome for *w*Wil (strain of *D. willistoni*) [141], assemblies published with

the first paper of this thesis, which includes three complete genomes of *w*Pau (strain of *D. paulistorum*), *w*Wil, *w*Pro (strain of *D. prosaltans*) and 11 draft genomes that also include *w*Tro and *w*Spt (strain of *D. septentriosaltans*), and finally, assemblies produced in the second paper of this thesis, that include 10 complete genomes and two draft assemblies from *w*Pau, *w*Wil and *w*Tro.

Chapter III: Mitonuclear Interactions

In Chapter II of this thesis, I discussed the obligate, long-term mutualistic relationships found in nature between bacterial endosymbionts and their animal hosts, particularly in insects. Mitochondria also originated from such an ancient endosymbiotic association between an α-proteobacterium and a single-celled organism, which may have been an archaeal host or a more complex proto-eukaryotic cell [142, 143]. One major consequence of this ancient obligate endosymbiosis was the transfer of genes from the endosymbiont, and ultimately the mitochondrion, to the nuclear genome of the last eukaryotic common ancestor (LECA). In most animals, around 1,100 mitochondrial genes are nuclear encoded (N-mt), while just 37 are encoded by the mitochondrial genome (mt) and code for 13 proteins, 22 tRNAs, and 2 rRNAs (Figure 6) [144, 145]. These 37 mt gene products are vital for functions for the host, such as oxidative phosphorylation (OXPHOS), and for the mitochondria, such as translation. The products from these 37 genes, non-coding RNA or proteins, must interact with some of the 1,100 N-mt proteins to create RNA-protein or protein-protein complexes that perform OXPHOS and mitochondrial translation, respectively. When these mitonuclear interactions are disrupted and organismal fitness is reduced, the resulting mismatches are referred to as mitonuclear incompatibilities. In chapter III, I introduce mitochondrial functions that could potentially be targets of mitonuclear incompatibilities. I then describe the expected fitness effects for mutations occurring in mitochondrial proteins under a mitonuclear co-evolution scenario. Finally, I examine the case of mitochondrial introgression and argue that to avoid any associated fitness costs either nuclear co-introgressions of N-mt genes or *de novo* compensatory mutations in N-mt genes, should occur.

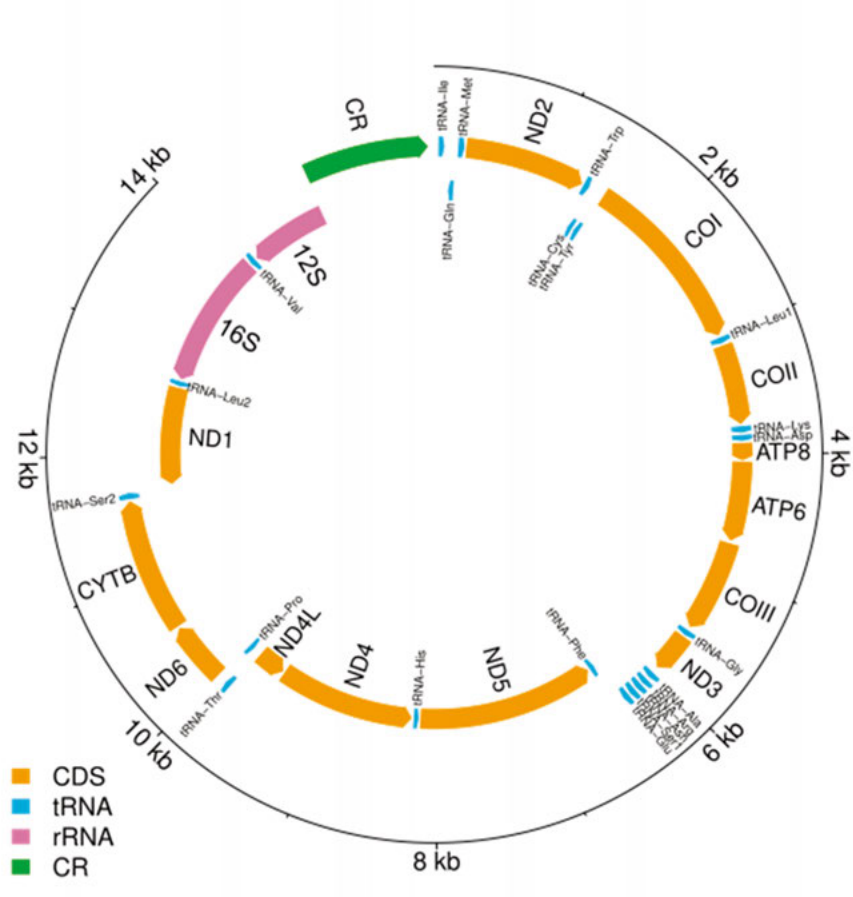

Figure 6. Mitochondrial genome of *D. paulistorum*, line C2 (Genbank ID: OP208743.1). The genome is 15,984 bases long and codes for 13 mitochondrial proteins for the OXPHOS complexes, 22 mitochondrial tRNAs and 2 ribosomal rRNAs. The genome is densely packed with genes, with the only non-genic region being the control region (CR), which takes part in mtDNA replication and translation.

Mitochondrial functions and mitonuclear interactions

Here I will introduce the mechanisms with which OXPHOS and mitochondrial translation are facilitated to highlight the stages at which mitochondrial incompatibilities may arise.

OXPHOS facilitates energy production for cells by extracting electrons from molecules produced by the tricarboxylic acid cycle (TCA) and transferring them, via the electron transport chain (ETC), to oxygen molecules [146]. The ETC consists of five protein complexes: the NADH dehydrogenase (complex I), the succinate dehydrogenase (complex II), the cytochrome bc_1

(complex III), the cytochrome c oxidase (complex IV) and the ATP synthase (complex V) (Figure 7). Complexes I and II transfer electrons from NADH and $FADH_2$, respectively, to ubiquinone (Q), creating ubiquinol (QH_2). Complex III transfers electrons from QH_2 to cytochrome c. Complex IV transfers electrons from reduced cytochrome c to oxygen, creating water. While complex I, III and IV catalyse electron transfer, they move protons from the mitochondrial matrix to the mitochondrial intramembrane space, thus creating a chemical gradient of protons, called protonmotive force. Finally, complex V uses the built up protonmotive force by allowing the passive flow of protons through it and back to the mitochondrial matrix while also converting ADP to ATP (Figure 7) [147].

In Eukaryotes, each OXPHOS complex consists of multiple proteins identified as "core" or "central" subunits if they have homologues in the bacterial ETC**,** and if not, they are called "accessory" or "supernumerary" subunits [148]. OXPHOS proteins may also have testis-specific paralogues caused by gene duplication and may serve functions other than OXPHOS [149, 150].

All OXPHOS complexes, except for complex II, include N-mt and mt proteins, which makes them potential sources of mitonuclear incompatibilities. Specifically, complex I includes 7 mt proteins (ND1-6, ND4L) out of 50 in total, in complex III cytochrome b is the only mt protein out of 14 total proteins, and complex IV includes three mt proteins (COI-III) out of 22 total proteins. Complex V is an F-type ATPase with two structural parts: the F0 with subunits B, C, D, E, F, CF6, OSCP, the mt ATP6 and ATP8 and the F1 structural part with subunits α, β, γ, δ, ε [151, 152].

Moreover, some mitochondrial OXPHOS proteins must interact with N-mt assembly factors in order to assemble into protein complexes [153]. For example, in mammalian complex I, recruitment of mt-ND1 into the protein complex requires interaction with NDUFAF3, 4 and 6. In complex III of yeast, insertion of mt-CYTB into the inner mitochondrial membrane requires interaction with UQCC1 and UQCC2 assembly factors. In complex IV, COX14-COA3 bind to mt-COX1 to create a module involved in complex assembly [153].

Other non-protein mitochondrial products that interact with N-mt proteins are the mitochondrial tRNAs and the two mitochondrial rRNAs 12S and 16S. Regarding tRNAs, in *D. melanogaster* each mitochondrial tRNA is charged with its respective amino acid by catalysis from its own aminoacyl tRNA synthetase (aaRS). The exception is glutamine, which does not have its own aaRS, but instead the tRNA of glutamine is charged with a glutamate amino acid via Glu-RS and then modified to glutamine via the amidotransferase Gat [154]. Regarding the 12S and 16S rRNAs, they interact with N-mt proteins for base modifications, for example methylation in 1369U and 1370G of human 16S

rRNA, and mitoribosome assembly, for example 12S rRNA interacting with ERAL1 for assembly of the small ribosomal subunit [155]. Thus, the interactions between mitochondrial noncoding RNAs and their N-mt proteins could be another source of mitonuclear incompatibility.

Finally, regulatory sequences like the origin of replication and promoters of the mitochondrial genome interact with N-mt proteins to perform replication and transcription of mt genes [156]. Disruption of those interactions will also cause mitonuclear incompatibilities.

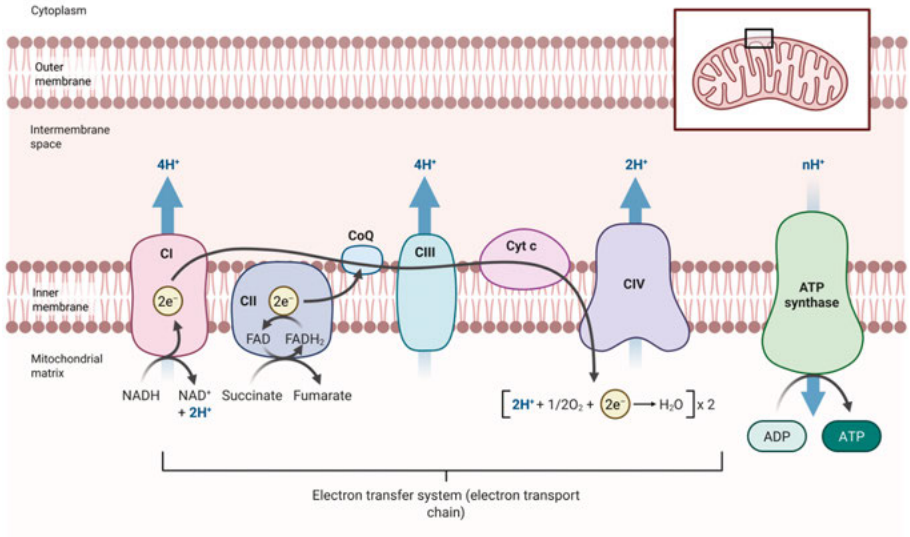

Figure 7. Electron transport chain. Five protein complexes facilitate OXPHOS in mitochondria. Electrons (e^-) are transferred from NADH and $FADH_2$ to oxygen, via complexes I-IV, while also transferring protons (H^+) from the mitochondrial matrix to the mitochondrial intermembrane space. This creates a chemical gradient, called protonmotive force, between the two sides of the mitochondrial inner membrane, which is used by the ATP synthase (complex V) to produce energy for the cell by converting ADP into ATP. *Created in BioRender. Papachristos, K. (2026) https://BioRender.com/e87u42i.*

Mitonuclear co-evolution

Mitochondrial functions are performed by protein-protein or RNA-protein complexes where mitochondrial RNA and proteins physically interact with N-mt proteins. These intergenomic interactions are epistatic, meaning that the phenotype/fitness of a mitochondrial locus *M* with alleles M_1 and M_2 is conditioned on the alleles that are present on another nuclear locus, *N* with alleles

N_1 and N_2 [157]. The sign of the epistasis between *M* and *N* loci points to how the nuclear and mitochondrial genomes co-evolve. For example, assume N_1 and M_1 are ancestral alleles and M_2 and N_2 emerged in one specific population. If genotypes that have interacted for a long time and emerged within the same population, i.e., M_1N_1, M_2N_1 and M_2N_2, have greater fitness than non-co-evolved combinations in mixed populations, i.e., M_1N_2, then mitonuclear interactions between the *M* and *N* loci are evolving mainly under purifying selection and only non-deleterious mutations that preserve compatibility between *M* and *N* can be fixed. Moreover, it is usually argued that mitochondrial genes are the ones driving mitonuclear co-evolution, as they tend to have a faster mutation rate compared to the nuclear genome, except for vascular plants, where the nuclear genome evolves faster than the mitochondrial [158]. However, it could also be that non co-evolved combinations of mitochondrial and nuclear genotypes produce individuals of equal or even greater than co-evolved combinations. Some studies have found substitution rates of mt and N-mt proteins to be about the same [159], or report about the same positively selected sites between interacting and non-interacting mt and N-mt proteins [160]. Thus, compensation may not always explain patterns of mitonuclear coevolution [157]. Finally, the epistatic effects between nuclear and mitochondrial loci may manifest in specific environmental conditions, such as temperature [161] and diet [162], complicating their identification.

Mitochondrial introgression

Species hybridization used to be considered “rare or absent in nature” [163], as ecological and temporal isolation of species and hybrid sterility were expected to be strong barriers. This view started to change by the 60s as more evidence of hybridisation in plants and animals started to show up [164]. Today we understand that hybridization is quite common in nature [165] with numerous examples from plants [166–169], insects [170], fish [171], birds [172], and mammals [173–175] including humans [176].

A consequence of species hybridisation may be mitochondrial introgression. Introgression refers to the transfer of genetic material between species by backcrossing with their hybrids [177]. Mitochondrial introgression refers to the transfer of mitochondria from one species to another by the same process.

As discussed above, depending on the nature of the interactions between N-mt and mt proteins, mitochondrial introgression will have one of three possible outcomes, increase fitness in the recipient species, have no effect on fitness, or reduce fitness [157, 178]. The case for the latter is intuitive as introgressed mitochondria interact with nuclear backgrounds with which they have

not co-evolved with, and it is supported by several observations reporting mitonuclear incompatibilities in hybrids within and between species [162, 179–190]. These mitonuclear incompatibilities, introduced by mitochondrial introgression, are resolved in one of two ways (Figure 8). The first is with co-introgression of nuclear components compatible with the introgressed mitochondrion in the recipient species. There are multiple instances of N-mt and mitochondrial co-introgressions described in the literature [184, 190–196]. The second way is with *de novo* mutations in the incompatible N-mt genes to rescue the interactions with mitochondrial components [190, 197, 198]. Admittedly, if mitonuclear incompatibilities are not identified in the more well studied and known OXPHOS and aaRS proteins, then demonstrating mitonuclear compensation by *de novo* mutations may be harder than co-introgression. This is because the former may require more high-throughput methods that use a lot of hybrid crosses to identify associations between N-mt and introgressed mitotypes that result in reduced hybrid fitness and are incompatible [190, 199]. In contrast, in the case of co-introgression, the source of mitonuclear incompatibilities is more likely to be identified in the co-introgressed N-mt genes [184, 195].

Another possibility for mitochondrial introgression is to increase fitness in the recipient species. Mitochondrial introgressions have been proposed to provide adaptation to environmental conditions [200–202], to replace degraded mitochondria due to Muller's ratchet [203].

Mitochondrial introgressions may also be driven by non-adaptive processes. The first are demographic processes of species. For example, in range expansion of an invading species into the range of a local species, mitochondria are expected [204] and have been observed [204, 205] to transfer from the local species to the invading. Another case of non-adaptive mitochondrial introgression is via drive mechanisms. Maternally inherited endosymbionts that manipulate host reproduction to spread into their populations will be co-inherited with specific mitotypes and as a side-effect, may lead to mitochondrial replacements [206–208].

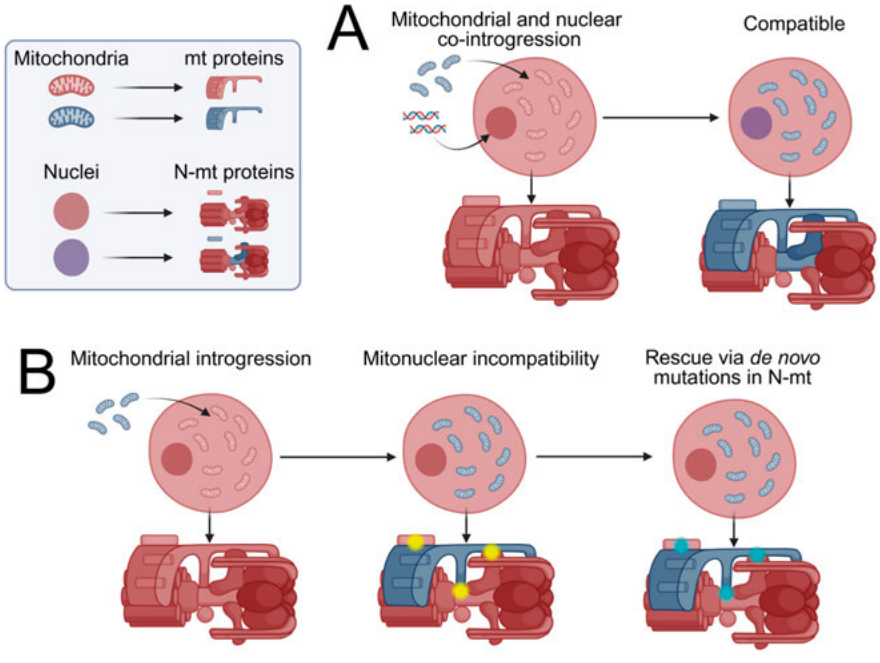

Figure 8. Possible paths for mitonuclear compatibility after mitochondrial introgression. A hypothetical mitochondrial protein complex comprises one mt protein and several N-mt proteins. Proteins that have been co-evolving together are shown in red and proteins introgressed from another species are shown in blue. Red nuclei represent the native nuclear background of the species, and purple nuclei the nuclear background with introgressed N-mt genes. A) Compatibility between the protein complex and the introgressed mt protein can be maintained if other N-mt proteins of the same nuclear background (blue) that properly interact with the mt protein are also co-introgressed with it. B) Introgression of mitochondria (blue) leads to poor interaction of that protein with the native N-mt proteins (red) at specific sites with which they interact (yellow glowing marks). If *de novo* mutations arise in those N-mt proteins at those specific sites (cyan glowing marks), they can rescue proper protein interaction. *Created in BioRender. Papachristos, K. (2026) https://BioRender.com/gctu6l7.*

Mitochondrial introgression can be inferred from discordance between nuclear and mitochondrial phylogenies, particularly when topological incongruence is accompanied by mitochondrial divergence or coalescence times that are unexpectedly recent relative to nuclear estimates of species divergence [209, 210]. Methods used for detecting introgression test against the null hypothesis of incomplete lineage sorting (ILS) accounting for the observed topological or temporal incongruencies [211]. ILS refers to ancestral polymorphism persisting after successive speciation events and being inherited to descendant species stochastically, thus potentially creating a phylogeny different than the species tree [212]. The framework used to model gene lineage coalescence within ancestral populations across a species tree is the multispecies

coalescence (MSC), and it allows for ILS by default [213]. If one examines coalescence of mitochondria in a simple speciation scenario of MSC with no gene flow between the species, the time of the mitochondrial ancestor of the two mitochondria between the species should be $T_{mtMRCA} = T_{divergence} + T_{coalescence}$, where $T_{divergence}$ is the time at which the species diverged, and $T_{coalescence}$ is the time span between mitochondria to coalescence and species divergence [214]. Under the MSC, $T_{coalescence}$ follows an exponential distribution, which can be used to calculate the probability of two lineages having coalesced after $T_{coalescence}$ generations [213]. Thus, by estimating species $T_{divergence}$ and T_{mtMRCA} and if $T_{mtMRCA} \geq T_{divergence}$, then one can estimate how unlikely it is for ILS to explain the time difference $T_{mtMRCA} - T_{divergence}$, or equivalently, how likely it is to be explained by mitochondrial introgression. If $T_{mtMRCA} << T_{divergence}$, then mitochondria must have introgressed.

Chapter IV: Nuclear mitochondrial insertions

The final paper included in this thesis focuses on a unique genomic element that was recently identified in the genome of *D. paulistorum* O11 line from the OR semispecies, called Mega-NUMT [215]. Montoliu and co-authors resolved 12 in tandem and near-complete nuclear mitochondrial insertions (NUMTs) of the mitochondrial genome at the end of a contig in their draft assembly. The set of these 12 near-complete in tandem NUMTs forms the Mega-NUMT. However, due to the repetitive nature of the Mega-NUMT, it is possible that not all copies are resolved in the assembly. Based on qPCR and sequence depth data there may be more unresolved NUMT copies in the O11 line, potentially up to 60. In the fourth paper of this thesis, we describe the Mega-NUMTs in more *D. paulistorum* lines from the AB, AM and CA semispecies and using comparative genomics we attempt to reconstruct the evolutionary history of these elements. This chapter discusses the characteristics of NUMTs, their significance in scientific research and medicine and possible ways they could form Mega-NUMTs.

NUMTs and NUMTogenesis

NUMTs are insertions of mitochondrial DNA sequences into the nuclear genome. The process of establishing *de novo* NUMTs is called NUMTogenesis. NUMTogenesis seems to be universal, as NUMTs have been detected across multiple eukaryotic genomes, from protists to humans [216, 217]. NUMTogenesis also seems to be an ongoing process through an organism's life, first because NUMTs are linked to aging [218] and non-inherited diseases such as cancer [219], and second, direct transfer of DNA from modified mitochondria to the nucleus of yeast has been demonstrated to occur at around 2×10^{-5} times per cell per generation [220]. NUMTogenesis also involves the germline, creating insertions that are heritable to the next generation. Population genomics in primates and humans shows intraspecies diversity in NUMTs [221, 222], and genome sequencing of mother-father-child trios reveals private NUMTs in children, suggesting a *de novo* insertion rate of $2.95\text{-}88.1\times10^{-5}$ per generation in the human germline [222]. Also, comparative genomics between

humans, Neanderthals and Denisovans has revealed NUMTs that are shared and are millions of years old [223].

NUMTogenesis is not a well understood process, and it is not known if it occurs in the same way across different organisms. In flowering plants, for example, given that LTR retrotransposons are abundant in their genomes [224], NUMTogenesis may involve transfer of NUMTs to the nucleus as cDNA via reverse transcription of mitochondrial mRNAs. One example in line with this hypothesis comes from papilionoid legumes, where a mitochondrial sequence of *cox2* was transferred into the nuclear genome of several species of the Phaseolinae (*Vigna radiata*, *Vigna angularis*, and *Phaseolus vulgaris),* and Glycininae (*Glycine max*) groups [225, 226]. This insertion became a functional N-mt gene [227], leading to the deletion of the mitochondrial *cox2* in Phaselolinae [228, 229], while in Glycininae the mitochondrial copy is retained but is silenced [230]. The *cox2* NUMT turns out to have inserted into the genome as cDNA after reverse transcription, based on greater similarity to edited mitochondrial mRNA than to mtDNA sequence [225]. However, other mechanisms directly involving mtDNA must also contribute to NUMTogenesis, as NUMTs may span multiple adjacent mitochondrial genes up to the complete mitochondrial genome. For example, in *Arabidopsis thaliana* there are large pericentric insertions comprising nearly the entire mitochondrial genome in the short arm of chromosome 2 [231], and in humans some NUMTs are as long as the complete mitochondrial genome [222]. Moreover, studies in yeast examining NUMTogenesis reject or consider unlikely the involvement of RNA intermediates in the process [232, 233].

For direct transfer of mtDNA into the nuclear genome, there should be at least three conditions fulfilled. First, mitochondrial DNA should become fragmented by double strand breaks (DSBs) [234]. Given that NUMTs are linked to disease and aging, the source of mtDNA fragments could be dysfunctional mitochondria due to environmental stress. Environmental factors such as radiation and heat stress, or free radicals that increase with age, will impair mitochondria by increasing DSBs [235].

Second, mtDNA fragments should escape the mitochondrion. Escape of mtDNA to the nucleus has been studied in several yeast mitochondrial escape (yme) mutants [236]. It seems that mtDNA escape to the nucleus is regulated by N-mt proteins via multiple pathways [232]. One such N-mt protein in yeast is Yme1, a subunit of a homo-hexameric i-AAA proteinase that is multifunctional and regulates mitochondrial protein homeostasis [237]. *yme1* mutants increase NUMT insertions via a vacuole-dependent pathway [238]. NUMTs also increase significantly in human cell lines with knockouts of *YME1L1*, the homologue of *yme1* [219]. Another protein identified to be regulating NUMT

insertions in yeast is Nuc1. Nuc1 is a nuclease that degrades long mtDNA fragments (>10 kb) that would escape to the nucleus [233]. It has also been shown that Nuc1 must be localised at the mitochondrial intermembrane space (IMS) for mtDNA degradation and that mitochondrial escape to the nucleus involves DNA degradation in vacuoles during mitophagy [239]. Yu and co-authors developed a high throughput amplicon sequencing method called Break-Ins to screen for short NUMT insertions in nuclear genomes via non-homologous end joining (NHEJ) [233]. Using this technique with already established assays in yeast to estimate transfer of mtDNA to the nucleus, they suggest that Nuc1 preserves nuclear genome stability by degrading retrotransposon cDNA and long mtDNA fragments. Thus, it seems that mtDNA escape to the nucleus is common and controlled against by mitonuclear interactions, which, if disrupted, may promote NUMTogenesis.

The third condition for NUMTogenesis, independent of RNA intermediates, is for the mtDNA fragments to enter the nucleus and to repair DSBs via NHEJ in the nuclear genome. Classical NHEJ repairs DSBs by directly ligating DNA ends with no homology or 1-4 bp of junctional microhomology [240]. Microhomology is observed in sites of NUMT insertions in yeast [234] and humans [241], supporting the role of NHEJ in NUMTogenesis.

NUMT insertion sites in a nuclear genome are expected to be biased against some genomic locations, as they could disrupt gene structure and their regulatory elements and be selected against. Indeed, NUMT distribution does not seem to be completely random. When examining NUMT patterns in humans, NUMTs are distributed across all chromosomes, with higher densities on some autosomes and a lower density on the X chromosome [222]. NUMTs are found near or within duplicated areas of the genome, but are underrepresented in regulatory elements, SINEs, simple repeats and gene introns, while NUMTs with the most frequency in human populations are found more than 500 bp away from the transcription start site of any gene [222]. Another study in humans also finds that fixed NUMTs, i.e., NUMTs shared by all members of a species, are enriched in intergenic regions and are underrepresented around TEs, while non-fixed NUMTs do not have a bias for absence of genes or TEs in their location [242]. Other studies in human, rhesus monkey, mouse and rat genomes, find that NUMTs are likely to be flanked by retrotransposons and AT oligomers and are more likely to be found in open chromatin regions [243]. Similar patterns are seen in yeast, where NUMT insertions seem to localise in non-genic areas and around LTR retrotransposons [234]. In honeybees [244], 84% of NUMTs are in non-genic regions, but there is no significant correlation between chromosome size and NUMT number, unlike in humans where this correlation is strong [222]. It thus

seems that NUMTs may insert preferentially around TEs which may have caused DSBs with their transposition, but NUMTs that are fixed across populations and species will probably be in less dynamic and less functionally important areas of the genome.

Finally, NUMTs could be selected against if they modify gene expression of nearby genes. Part of the D-loop (a.k.a. mitochondrial control region) of mitochondrial genomes seems to be missing from NUMTs in humans [245] and other mammal species [243]. This was attributed to the D-loop's fast rate of evolution making it undetectable, or to mitochondrial transcription factor A (TFAM) interacting with the D-loop and protecting it from fragmentation [246] and insertion in the nucleus as a NUMT. In a recent study in humans, the 3-prime area of the D-loop, mtDL3, is deleted in fixed NUMTs in humans and other primates, but not in non-fixed NUMTs [242]. The authors also demonstrated promoter activity of mtDL3 by expressing GFP in human cells. These results suggest that recent and non-fixed NUMT insertions with the mtDL3 region may affect gene expression in nearby genes and thus be under negative selection in primates.

NUMTs can cause problems in scientific research and medicine. Mitochondrial DNA has been used extensively to identify known species and to identify cryptic species, i.e., species that cannot be distinguished morphologically from others [247]. Thus, older NUMTs could be mistaken for DNA from diverged mitochondrial genomes, which can potentially lead to species misidentification or description of non-existing cryptic species [248]. An example of the latter is demonstrated by studies in the leatherback turtle *Dermochelys coriacea*, in which NUMTs were mistaken for mitochondrial DNA, leading to the conclusion that two species or subspecies of turtles exist and that they diverged 13.5 million years ago [248]. NUMTs can also appear as instances of heteroplasmy in PCR and pedigree data, falsely suggesting paternal inheritance of mitochondria [249, 250]. Moreover, multiple diseases have been associated with specific mitochondrial mitotypes [251, 252], and their detection in clinical genetic testing may be complicated by the presence of NUMTs. Therefore, accurate characterization of NUMTs in genomes is essential for ensuring reliable scientific and clinical interpretations.

From NUMTs to Mega-NUMT

After their insertion in the nuclear genome, NUMTs may form more complex patterns. For example, in domestic cats, a mitochondrial segment has been amplified 38-76 times as a tandem repeat macrosatellite, and based on phylogeny, this NUMT insertion seems to have happened in the ancestor of four

species of the genus *Felis*, around 1.8-2.0 million years ago [253]. In chromosome 2 of *Arabidopsis thaliana*, a NUMT insertion of around 618 Mbp of the complete mitochondrial genome with two internal tandem duplications was reported [231]. In humans, complex NUMT insertions have been described, where NUMTs are part of two different fragments and of different strands of the mitochondrial genome [250]. In another study, multiple NUMT insertions were estimated to range from 700 to 900 kbp and comprise 45-56 tandem copies of the U4c1 mitotype, though the insertion has not been sequenced with long reads [254]. Another insertion in humans 127.8 kbp long consisted of several mitochondrial fragments from various positions and strands of the mitochondrial genome [242]. Such multicopy mitochondrial DNA insertions in the nuclear genome are called *Mega-NUMTs* [249].

Mega-NUMTs may be a result of inherent genome mutation rates, rather than NUMTogenesis. Once NUMTs are inserted in the nucleus, they have no mechanism for replicating and increasing their number in the genome, except for processes that affect the specific genomic areas in which they are inserted. For example, genomic areas rich in tandem repeats such as satellite DNA are more prone to tandem duplication and deletions via strand slippage, and areas rich in interspersed repeats like TEs are more prone to larger segmental duplications and deletions via unequal crossing over [255, 256]. Thus, a few NUMTs inserted in a genomic area with multiple TEs or tandem repeats could then increase in copy number because of the high rate of segmental duplications in that genomic area.

Chapter V: Sequencing and Genomics

Most of the work for this thesis involved genomics and bioinformatics to understand evolutionary relationships of the Neotropical *Drosophila*, their mitochondria and their *Wolbachia*. Because the genomes of those three differ, they present different challenges. Thus, in this chapter I first discuss the genome architecture in prokaryotes and eukaryotes to highlight their particularities when it comes to work with genomics in those systems. Then I provide a brief history of DNA sequencing technologies to this day, followed by an introduction to genome assembly and annotation. Finally, I discuss how comparative genomics can help us investigate the evolution of organisms. I close this chapter by describing a few examples from the work of this thesis, where the assembly process and other downstream analyses would require troubleshooting, either because of real biological differences or differences in data quality between samples.

Genome architecture of prokaryotes vs eukaryotes

Genomes can be categorised into two types based on their content and organisation, the genomes of prokaryotic and the genomes of eukaryotic organisms. As a general rule, prokaryotic genomes are smaller than eukaryotic, usually around 2-5 megabases long [257], and comprise one circular chromosome and potentially other smaller circular DNA molecules, called plasmids [258]. In contrast, eukaryotic genomes range from some megabases to hundreds of gigabases [259] and have multiple linear chromosomes per haploid genome [258]. Some exceptions do exist, with some prokaryotes found to have more than one chromosome [260–262] or linear chromosomes [263], and some eukaryotic genomes being smaller than some prokaryotic [264]. Prokaryotic genomes are more gene-dense than eukaryotes, with the gene density of the former being around 1 per 1kb [265], while in the latter being orders of magnitude lower [259]. Genes in prokaryotes are uninterrupted, while in eukaryotes they are usually organised in alternating areas of coding and non-coding DNA, called exons and introns respectively. Moreover, alternative splicing of exons in eukaryotic transcripts allows for multiple gene products [266]. Also, many genes in prokaryotes form groups of co-transcribed genes called operons,

while this is rare in eukaryotes [267]. Prokaryotic genomes have typically no or few transposable elements per megabase [268], as opposed to eukaryotes where they may comprise 14-75% of the total genome [224, 259]. Simple repeats are also rare in prokaryotes [269] but quite common in eukaryotes [270], some of which comprise the centromere of each chromosome. In contrast, prokaryotic chromosomes do not have centromeres [258]. In conclusion, prokaryotic genomes are generally smaller, more gene-compact with a simpler organisation and lack repeats and non-coding DNA, while eukaryotic genomes are orders of magnitude larger, with a more complicated gene organisation, less gene-dense and more repeated.

Because of these differences, it is more difficult to assemble [271] and annotate [272] eukaryotic genomes than prokaryotic. A quick search for bacterial and eukaryotic genomes available in NCBI datasets [273] at the time of writing of this thesis (August 2026), shows more total submissions of assemblies in bacteria (3,244,712) than eukaryotes (67,157). Of those 3% (97,108/3,244,712) is at chromosomal or complete level for bacteria, compared to 26% (17,585/67,157) in eukaryotes, and from those, 77% (75,132/97,108) and 7.8% (1,363/17,585) are annotated, for bacteria and eukaryotes respectively. So even today, the amount of high-quality and annotated eukaryotic reference genomes lags behind bacteria.

Sequencing technologies

DNA sequencing started with the Sanger method in 1977 [274], which was used to sequence the first human genome in 2001 [275]. The strategy followed in the human genome project was the hierarchical shotgun sequencing of large parts of the human genome, or *inserts*, that were cloned into bacterial artificial chromosomes (BACs). First, a physical map of overlapping inserts from many BACs was constructed, and then each BAC insert was fragmented into smaller pieces of DNA and sequenced via the Sanger method. Sequencing became more efficient with the introduction of the second generation of sequencing technologies, referred to as next generation sequencing (NGS) [276]. NGS follows a whole genome shotgun (WGS) sequencing approach which skips the tedious work of establishing BAC clones and their physical mapping. Instead, in WGS the genome is randomly fragmented to create inserts, and DNA adapters are in vitro ligated on the inserts, at a stage called library preparation. The adapters allow the amplification of the inserts, the pieces of DNA that will be sequenced, resulting in pools that are later sequenced by various sequencing technologies. The result of the process is sequences from multiple inserts, which are called *raw reads*, or just *reads*.

The output of NGS technologies is short raw reads. Depending on the technology, read length may range from 50 to 300 bases [277]. The most common short read sequencing technology today is that from Illumina which may produce reads from 100 to 300 bases depending on the service [277], with an insert size between 350 and 550 bases [278]. Illumina uses a "sequencing by synthesis" approach, where a DNA polymerase incorporates modified nucleotides into a growing DNA strand complementary to the DNA template that is to be sequenced [276]. Sequencing can be done in single end (SE) mode by adding only one adapter sequence during the library preparation and thus each insert is sequenced from one end only. The alternative is paired end (PE) mode, where two adapters, one at each end of the insert, are added and each insert is sequenced from both ends [277]. Thus, reads in PE that originate from the same insert and are linked. This linkage information makes PE sequencing useful for identifying structural variants or the chromatin organisation of a reference genome [279]. This is because inserts have their lengths distributed around an expected size that is set during the library preparation by randomly fragmenting the DNA [278], making PE reads to have an expected "insert distance" from each other. PE reads mapped on a reference genome that do not have the correct orientation, or the expected insert distance will be indicative of genomic variants such as inversions and translocations compared to a reference genome. Moreover, paired reads that exceed half the insert length will overlap with each other. Overlapping paired reads can be merged into one read and may increase sequencing accuracy [280].

NGS sequencing with short reads revolutionised the field of genomics by allowing high throughput of genomic data. By 2020, there had been more than 400,000 bacterial genomes deposited in RefSeq and more than 80% of those were sequenced by Illumina [281]. But as mentioned above, bacterial genomes being smaller and less repetitive, have been easier to assemble and annotate compared to eukaryotic genomes. When it comes to the assembly process, short reads from repeated regions that are larger than the read length cannot be uniquely assigned to their correct genomic location [282]. This is because assembly algorithms place reads in graphs and infer the original sequence by finding the best graph path [283]. The problem is that the resulting graph paths are ambiguous and therefore create gaps in the assembly [284]. Thus, the larger and more repeat rich genomes of eukaryotes have been more difficult to assemble completely with short reads.

More recent sequencing technologies, referred to as third generation sequencing (TGS), aim at solving this issue, by sequencing inserts of larger size. There are two long-read sequencing technologies available today offered by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) [277].

Both produce raw reads tens of kilobases in length, allowing for a greater resolution of repetitive regions and large structural variation in genomes. They can also avoid the need for PCR amplification in library preparation and the creation of clonal populations of DNA fragments during sequencing [276]. When they first launched, both sequencing technologies produced long reads with high error rates (around 15% in PacBio and 30% in ONT) compared to Illumina short reads (around 0.1% with Illumina HiSeq 2000) [285], but improvements in the technologies, like the use of circular templates in PacBio HiFi sequencing, or improvements in the chemistry and base-calling algorithms of ONT, have now made the error almost equal [283].

Both NGS and TGS have systematic biases and random errors. Regarding systematic biases, Illumina sequencing is biased against extreme GC content in DNA fragments [286, 287]. ONT may incorrectly infer the length of homopolymers [288]. Both PacBio and ONT have more difficulty in sequencing short satellite DNA compared to other regions with higher sequence complexity [289] and library preparation protocols may still introduce biases [290]. It is important to be aware of such biases and account for them if possible. Random errors, on the other hand, can be reduced by sequencing with high depth, i.e., the average number of times each nucleotide position in a genome is sequenced. However, high depth will not remove systematic errors.

An overview of genome assembly and annotation

As already mentioned, the first efforts to sequence the human genome relied on physical maps of the BAC inserts to be sequenced. It is uncommon to use BACs in today's genome projects [291], as a lot of work is required to create them and make physical maps for them [292]. Instead, the most effective strategy used today involves extracting whole genome DNA and sequencing with long read technologies as well as with short reads that may be used to correct assembly errors or provide linkage information [293]. Then, assembly algorithms piece together the unordered reads by using unique points of overlap among them and creating graphs connecting the reads. Two types of assembly algorithms exist. The de Bruijn graph (DBG) algorithm uses unique kmers, substrings of the reads, to create graphs and find the best way to piece them together [294]. The overlap-layout-consensus (OLC) algorithm uses all-vs-all pairwise alignments of reads to find overlaps of minimizers, a subset of all possible kmers in each read, and makes a layout using graphs and infer a consensus sequence [295]. The output of the assemblers is *contigs*. Contigs are contiguous DNA sequences assembled from overlapping sequencing reads without any gaps [296]. They are the result of genome assembly and represent

regions of the genome where the order and orientation of the assembled bases are known with confidence based solely on the sequence data. Today, with the help of long reads, a contig may represent a complete chromosome, but in order to increase overall contiguity of an assembly, it is common to connect contigs based on linkage information from PE reads, long reads, or chromatin conformation capture (Hi-C) data [297]. The result of linking contigs based on linkage information is s*caffolds*. In scaffolds, adjacent contigs are separated by gaps of unknown length and sequence.

Assemblies should be considered as hypotheses of what the genome of a species may look like, rather than its true representation [296]. This is because of multiple reasons. First, genome assemblies are inferred by heuristic algorithms with limitations [298, 299], and thus contigs and scaffolds are reconstructed by sequencing reads and linking information, rather than being directly observed. Second, diploid and polyploid organisms will have different haplotypes with various SNPs and structural variants. Areas of the genome with high heterozygosity will thus be assembled in separate contigs, effectively creating duplications in the assembly, while homozygous regions are collapsed into one contig [300, 301]. This could be mitigated by read-based haplotype phasing [302], the process in which reads are assigned to the haplotype they are inherited from based on SNPs. Moreover, it is common for DNA to be extracted by multiple different individuals for small sized species, as sequencing, especially with long reads, requires a lot of DNA [303]. Thus, multiple haplotypes may be sequenced in a pooled sample, creating the same problems in assembly as discussed. One could overcome this by sequencing individuals from populations kept for a long time in the lab, as they should be subjected to high inbreeding and drift, which reduces heterozygosity [304]. Obtaining a reference genome from an inbred population that is thought to be representative for most of the species may be preferable, as it would reduce assembly complexity.

Once an assembly of a genome is complete, the next step is to annotate it. Common features to annotate in both prokaryotic and eukaryotic genomes are genes, pseudogenes and repeats [305, 306]. Genome annotations for model organisms will provide more information on regulatory sequences, expression data, metabolic pathways, protein 3D structure, population variants, phenotype information, protein clustering groups, the position of centromeres and telomeres, CpG islands etc [307–309].

Gene annotation is done in two stages, the first being *structural annotation*, where the position of genes and their structure is identified, and the second being *functional annotation*, where information about the genes is retrieved, such as gene names, protein functions, cellular location, etc [305]. Structural annotation

tools take two approaches in gene search. The first is *ab initio* search. Such tools are trained on organism specific genomic features, like codon frequencies, GC content, length of introns and exons, to identify genes [310–312]. They are fast to run, but are limited by being trained on specific organisms, and cannot report isoforms or untranslated regions (UTRs). The second approach is evidence-based search in which genes are identified via RNA alignments from the same or closely related species or protein alignments [313, 314]. This approach requires RNA data which may require a lot of effort to produce, protein annotations from other relatively closely related organisms, which may not be possible in for larger and rare taxonomic groups. In practice, structural annotation is done by running pipelines that use multiple software. Traditionally, annotation pipelines for prokaryotes involves *ab initio* programs and annotation transfer from proteins matching to the predicted genes [306, 315, 316], and this is enough to identify gene structure of the uninterrupted, single-CDS genes of bacteria, though some pipelines might also incorporate extrinsic evidence for gene structure, like RNA and protein alignments [317]. In contrast, structural annotation pipelines for eukaryotes, combine both *ab initio* and evidence-based approaches [318, 319] to account for their larger and sparser exon-intron organised genes. Combining both approaches increases sensitivity, i.e., true positive gene calls, and specificity, i.e., reduce false positive gene calls [320]. Moreover, eukaryotic genome annotation is routinely preceded by repeat identification to increase annotation accuracy [321] as both in tandem [322] and interspersed repeats [323] affect gene prediction.

Functional annotation of both prokaryotes and eukaryotes comes with challenges and is an active research topic within the bioinformatics community [324].

The first challenge is that assigned functions of model organisms are passed as annotations to proteins with high sequence similarity, indicating homology [306]. Proteins are homologous when they share common ancestry [325]. Searching for homologous proteins based on sequence similarity can be problematic because similarity searches are susceptible to statistical and compositional biases [326], and sensitivity will drop when looking for homology in more distantly related species [327]. Ideally, function should be assigned between orthologues [326], i.e., homologous proteins that are a result of speciation events, rather than paralogues, i.e., duplicated proteins in the same organism. This is because paralogues may be assumed to persist only if they provide a new useful function [326], possibly due to neofunctionalization [325]. Moreover, assignment of function due to similarity is essentially a prediction of function, rather than proof, which would entail experimentally verifying the prediction. Few studies try to quantify misannotations in databases, [328,

329], but it seems that the rate may be increasing with the submission of new genomic data and errors may propagate from database to database [328].

A second challenge for functional annotation is that most of them have incomplete coverage. Incomplete coverage means that several predicted proteins have unassigned functions, either because there is no homology to any other known protein or because the function is predicted and not experimentally verified. Even in well studied model organisms annotation coverage is lacking to various degrees. In humans, around 8% (1,396/18,194) of the total experimentally verified proteins is of unknown function and around 21% (3,814/18,194) has its function predicted and remains to be characterised experimentally [330]. In *Drosophila melanogaster*, around 42% of the total genes have been associated with 10 or less publications and await manual curation and functional characterization [331]. In *Arabidopsis thaliana*, the model organism of plants and best annotated among them, around 15% (4,070/27,657) of all proteins are poorly characterised [332]. Even in *Escherichia coli* K-12 more than 2% of the proteome lacks sequence homology and 31.8% have predicted functions and remain to be experimentally characterised [333]. A broader evaluation of genome annotation coverage across the bacterial tree of life finds that 21-48% of the proteome of any bacterial species may have unknown function, in some taxa being as high as 86% [334]. Another problem is that there are only a few model organisms per broad taxonomic group. For example, annotation of bacterial genomes with experimentally supported gene ontology terms, can be done only from *E. coli* terms [324]. Given the substantial evolutionary distances across the bacterial domain, a single model species cannot adequately represent the functional diversity of all bacterial lineages, which contributes to the low coverage of annotation for non-model organisms [334].

In conclusion, genome annotations face many challenges in both the structural and the functional part, and thus they should ideally be manually curated, supported by experimental data, and updated frequently with new information. This may not be possible for most non-model organisms, but their annotations will improve with the establishment of more model organisms and improvement of existing ones.

Comparative genomics

The massive amount of genomic data accumulated until today have allowed researchers to approach questions regarding evolution of species, genome function and human disease in a comparative framework [335]. This framework of comparing genomes within or between species to answer such questions is called *comparative genomics*.

All living organisms descend from a common ancestor from which they evolved [336–338], and like so all genomes are a product of evolution. All four evolutionary forces, i.e., mutation, selection, genetic drift and gene flow [339], shape the content, the structure and the function of genomes [340]. Thus, by comparing genomes of organisms from different populations or species and by looking at the genomic differences separating them, we can infer how these evolutionary forces affected the ancestors of these organisms in the past. For example, Kosiol and co-authors compared the proteomes of six mammals to infer positive selection throughout mammalian evolution [341]. Another study evaluated germline mutation rates by comparing genomes from parent-offspring trios across vertebrates and identified life-history traits that affect variation in mutation rates [342]. In Neotropical butterflies, genome differentiation between three closely related species demonstrates past gene flow and identify genomic regions involved in adaptive introgression [343].

Another possibility is to link genomic differences to phenotypic differences and hypothesize about the function of unknown genes [344]. For example, the CI inducing proteins of *Wolbachia*, CifA and CifB, were first identified in the ovaries and sperm of *Culex pipiens* mosquitoes using proteomics methods [345] but became candidates for CI once it was realised that the *cifA* and *cifB* genes can be found in genomes of CI-inducing *Wolbachia* strains of arthropods, but not in non-CI-inducing strains [108, 109]. Later these genes were validated to induce CI [108, 109]. In another comparative study on of Darwin's finches, genomes from several species with different beak morphology were compared and identified variation in the *ALX1* gene to drive evolution of beak morphology in those species [346].

Moreover, comparative genomics can reveal information about the function of genomic areas and assist in protein annotation [344]. When comparing the genomes of distantly related organisms, identifying conserved areas suggests they evolve under purifying selection and are functional. For example, the conserved genomic areas between humans and dogs account for 5.3% of the human genome and some of these may account for conserved regulatory sequences controlling developmental genes [347]. In a study comparing genomes between three species of *Saccharomyces*, the genome of *S. cerevisiae* was used to identify conserved protein genes and non-coding elements in those the species and helped to improve the annotation of their genomes [348].

Homology, and more specifically, *orthology*, is an essential part of comparative genomics, to be able to compare equivalent genomic features between genomes and thus make informative comparisons [349]. I briefly mentioned the problems of inferring homology from protein sequence similarity when discussed functional annotation in the previous section. There are other methods

complementary to sequence similarity for inferring orthology in comparative genomics, such as protein clustering, synteny, and conservation of protein structure. For protein clustering according to OrthoMCL [350], proteins between multiple genomes are compared based on sequence similarity with all-against-all BLASTP comparisons, and proteins with the best reciprocal hits are clustered into groups of orthologues and recent paralogues. The strength of this method is that information across all proteins and species studied is considered to assign orthology, but its weakness is that it will underperform when provided incomplete proteomes, and therefore, having complete annotated genomes is important. Synteny can also help identify orthologues by comparing the orientation and order of genes between genomes [351]. This is usually done by first identifying all conserved genes or kmers (anchors) across all contigs or scaffolds and across all genomes to reconstruct synteny blocks based on dynamic programming or graph-based approaches [352]. As an example of this, synteny comparisons between humans and mice increased orthologous genes identified by 20% compared to sequence similarity methods [351]. Finally, orthologous proteins tend to have more similar structures compared to paralogues with the same level of sequence similarity, and orthologues are even more conserved on sites that bind ligands [353]. Thus, if 3D structures are available, or if predicting such is reliable, this can help in orthologue assignment.

Much of the comparative genomic work presented in this thesis relies on the described methods of sequence similarity, protein clustering and synteny and enabled many downstream evolutionary analyses. For example, they enabled phylogenetic inference for investigating evolutionary relationships among taxa (Paper I, II, III, IV), inference of historical introgression events (Paper III), detection of signatures of selection (Paper III), ancestral sequence reconstruction (Paper III), identification of gene content differences among genomes (Paper I, II, III), finetuning of OXPHOS and aaRS orthogroups (Paper III), and the identification of Mega-NUMTs present in the *Drosophila paulistorum* semispecies (Paper IV).

Some encountered methodological challenges

During the time we were developing the bioinformatic methods used in the papers of this thesis, many challenges occasionally presented themselves, requiring certain samples to be treated differently, or pipelines to be changed completely. In this section I will mention three of the most noteworthy instances, explain what the issues were, and how they were resolved.

The first came up in analyses for paper II. The *Wolbachia* strain of *D. paulistorum* MS, *w*PauMS, is infecting its host in a low-titer infection. As a result,

Wolbachia reads from this sample, were scarce and noisier compared to samples from high-titer infections. This meant that, first, to assemble a high completeness *Wolbachia* genome, we had to use multiple sequence data from the same sample, reaching a total of 140X in sequence depth for the host, just to recover enough *Wolbachia* reads. Second, the kmer size for read overlap during assembly, had to be substantially reduced compared to when high-titer samples were assembled. Third, because of the low-quality of *Wolbachia* reads in the MS sample, more read filtering was required. We removed all reads that had more than 50% of their total bases as soft or hard clipped after mapping on *Wolbachia* reference genomes. This step alone reduced the number of reads from 1,200X to 44X depth for a 1.2Mb *Wolbachia* genome.

The second instance concerned paper III and the assembly of some *D. paulistorum* lines that were recently collected and sequenced with PacBio long reads. Pools of 20 male and 20 female individuals were sequenced for all lines. Old lab lines, such as Prune, would have no different haplotypes and sequence data assembled high completeness genomes at the expected genome size. However, the recently collected lines had multiple different haplotypes between the 40 individuals in each sample, thus creating assemblies with high duplication percentages in BUSCO genes [354] and double or triple the expected genome size. Kmer analysis of some of those assemblies with Merqury [355], for example, line CR97, showed that some contigs represented more than four haplotypes. After trying multiple ways to clean the assemblies from the extra haplotypes, we found that the best results, i.e., high completeness, low BUSCO duplication and a genome size around 250 Mb, were after filtering out contigs with low coverage and manually removing any remaining large contigs that mapped more than once on the *D. paulistorum* L12 genome [40].

The third example, again from paper III, relates to the protein clustering of the Neotropical *Drosophila* proteomes. A few genes in some genomes could have different isoforms annotated compared to others. In most cases this would not be a problem, as at first, we chose to cluster the longest isoform from all genes for our comparative analyses. However, some of those genes that had different number of isoforms, would have a shorter isoform annotated as their longest when compared to other samples, and thus, those protein clusters would have some proteins with some exons missing compared to other orthologues. Manual inspection of the alignments of those protein clusters showed similar sequences, except for gaps in the places of the missing exons. However, these alignments would also come up as positive hits for introgression in some analyses. When replacing the sequences in the clusters of those proteins, so that all proteins had the same isoform, the positive introgression results went away. So, to address these cases, we

reclustered the proteomes with all isoforms included, and then pruned the resulting clusters to keep the isoforms with the shortest pairwise phylogenetic distance.

With these examples I want to highlight the difficulties in genomic bioinformatic research. First, solutions are occasionally case-specific. As I pointed out in the case of *w*PauMS, genome assembly of the same organism may not perform equally well across samples and this will depend on their properties. Second, progress is usually iterative. Choosing the most appropriate method for an analysis often requires multiple attempts and repeated evaluation of the output, as was the case when assembling PacBio reads from the recently collected fly lines. Unexpected results in downstream analyses may also come from inconsistencies introduced at an earlier step and therefore require revisiting and correcting previous parts of the pipeline, as I pointed out by the protein clustering and introgression analyses example. Finally, manual inspection of outputs and quality control are very important to have, to detect inconsistencies and errors that may otherwise lead to misleading results.

Research Aims

The Neotropical fruit flies of the willistoni and saltans groups, and specifically *D. paulistorum*, are a great system to study symbiosis and evolution. Four key features support their suitability for such studies. The first concerns their relevance to both symbiosis and speciation, which is that *D. paulistorum* is undergoing speciation and its semispecies are partially reproductively isolated. *D. paulistorum* would be a great case where bioinformatics and comparative genomics can be used to understand the ongoing genomic processes that induce reproductive isolation in its six semispecies. Moreover, reproductive isolation between the semispecies can partly be attributed to their *Wolbachia* endosymbionts, making the case for microbe-driven speciation [29].

The second feature of the Neotropical fruit fly system, highlights symbiosis and evolution from the endosymbiont perspective. *Wolbachia* strains infecting the willistoni and saltans groups are closely related to each other and seem to have lost the ability to induce CI in their host [356]. Thus, it is unclear how these facultative *Wolbachia* infections persist in these fruit flies. Additionally, *Wolbachia* in *D. paulistorum* is an outlier in that it is an obligate infection for its host [29, 137]. This provides a rare opportunity to compare the genomes of closely related endosymbionts that follow different lifestyles. Given the lack of obvious parasitism in the willistoni and saltans groups and the obligate relationship of *D. paulistorum* with *Wolbachia*, one could also examine whether the Neotropical fruit flies and *Wolbachia* have an older more intimate relationship, implying co-evolution between the two partners, or whether this is another case of transient *Wolbachia* infections. These points are discussed in paper I and II of this thesis, where in the former, co-evolution is examined within the willistoni and saltans groups, and in the latter, co-divergence is examined specifically within *D. paulistorum*. Thus, in papers I and II we address the following research questions.

Paper I:

- What is the history of *Wolbachia* infection in the willistoni and saltans groups? Did *Wolbachia* horizontally transfer between the species, or did the ancestor of the species become infected and *Wolbachia* has been co-speciating for millions of years with its hosts?

- What are the changes in genome structure and content that occur in endosymbionts during the early stages of host dependence increase?

Paper II:

- What is the relationship between *D. paulistorum* semispecies and their *w*Pau *Wolbachia* strains? Given the obligate relationship of all semispecies to their endosymbiont, did *w*Pau spread into the semispecies, or did it co-speciate with them?
- Are there genomic differences between *w*Pau strains from different *D. paulistorum* semispecies? If so, could these differences make the strains incompatible with each other and contribute to the reproductive isolation of the semispecies?

The third feature of the Neotropical fruit fly system that makes it interesting for studying is the complex evolutionary patterns in the willistoni subgroup, and specifically in *D. paulistorum*. In the past, mitochondrial introgression events have occurred within the willistoni subgroup. Baião and co-authors found that the mitochondrial and nuclear trees of the species in the willistoni subgroup are incongruent [38]. Specifically, for *D. paulistorum*, mitochondria are polyphyletic, discerning two mitotypes, α and β. Baião et al., 2023 argue that both mitochondria are introgressed into *D. paulistorum*. As explained in Chapter III, mitochondrial genomes produce proteins that interact with other N-mt proteins to perform vital mitochondrial functions, such as OXPHOS, and the expectation is that once mitochondria introgress into a new nuclear background, these interactions will be non-optimal and cause mitonuclear incompatibilities that reduce fitness. In such a scenario, mitonuclear compatibility can ensue, either by co-introgression of the compatible nuclear components with the mitochondria, or by *de novo* compensating mutations in the incompatible nuclear components. Therefore, in paper III, we ask the following.

Paper III:

- What mitochondria are more likely to have introgressed within the willistoni subgroup?
- Are there signs of mitonuclear incompatibility in the nuclear genome of *D. paulistorum* due to mitochondrial introgression? If so, is incompatibility resolved via nuclear co-introgression or via *de novo* mutations?

The fourth feature pointing out Neotropical *Drosophila* as a unique system to study genome evolution is the presence of a unique genomic element in *D. paulistorum*, a Mega-NUMT. Montoliu-Nerin and co-authors have demonstrated the presence of a repeated genomic element made of multiple in tandem copies of complete mitochondrial genomes, in the genome of *D. paulistorum*, line O11 of the OR semispecies [215]. There are 12 mitochondrial copies resolved in the Mega-NUMT of O11, but based on qPCR and read depth, and the fact that the element is at the end of a contig, there may be more copies of the mitochondrial genome. Additionally, the mitochondrial insertions of the Mega-NUMT are closely related to the α mitotype of *D. paulistorum*, while the O11 lines carry the β mitotype, hinting that α mitotype mitochondria was present in this line in the past. In paper IV we investigate more *D. paulistorum* genomes for Mega-NUMTs and address the following questions.

Paper IV:

- How are Mega-NUMTs organised in the genomes of the different semispecies of *D. paulistorum*?
- What is the evolutionary relationship between NUMTs from different semispecies of *D. paulistorum*?
- Are there any hints in the genome regarding the mechanism of Mega-NUMT establishment?

Summary of research papers

Paper I

In Paper I, we focus on *Wolbachia* from the willistoni and saltans groups. First, we identify the phylogenetic relationships of the sequenced strains to each other by using single-copy genes. *Wolbachia* from willistoni and saltans groups of fruit flies, as well as from *D. borealis*, *D. innubila* and *D. incompta*, form one monophyletic clade and we propose to refer to them as *w*Au-likes. The *w*Au-like strains are very similar to each other, with only 0.03% of the total 562,755 nucleic sites being parsimoniously informative. We looked for co-phylogenetic patterns between *Wolbachia* and fruit flies of the willistoni and saltans groups and find that the phylogeny of the endosymbionts is congruent with the mitochondrial, but not the nuclear phylogeny of the hosts. The nuclear host phylogeny suggests that *w*Pau, the obligate strain infecting *D. paulistorum*, has been horizontally transmitted from another host. If *w*Pau and *D. paulistorum* are not considered, *Wolbachia* and host phylogenies are congruent, implying co-speciation in the willistoni and saltans groups. However, the very short branch lengths of the *Wolbachia* phylogeny, are problematic, as under co-speciation they imply an extremely slow rate of evolution, on the order of 10^{-11} substitutions per site per year, or 10^{-10} substitutions per synonymous site per year. The latter rate is about one order of magnitude lower than rates where *Wolbachia* is claimed to co-speciate with its hosts in *Nasonia* and *Nomada*. Evolutionary rates or generation time for *Wolbachia* are not known, so if these rates are an underestimation, then the temporal component for co-speciation between the *w*Au-likes and the Neotropical flies does not stand, and the endosymbionts must have been horizontally transmitted. Next, we evaluate the genome content of complete *Wolbachia* genomes and see that *w*PauO11 has an expansion of IS4 elements, accounting for about 10% of the total genome size, while this expansion is not observed in related genomes. Moreover, the genome of *w*PauO11 has more genomic rearrangements, more pseudogenes and fewer WO prophage sequences, compared to related genomes. While WO prophage DNA is being reduced by *w*PauO11, a specific WO region of 11 genes, called Undecim, is duplicated, suggesting an important role in *w*PauO11 and *D. paulistorum* symbiosis. We speculate that the observed genomic differences between *w*PauO11 and other *w*Au-like strains are a consequence of *w*PauO11 transitioning from facultative to obligate for *D. paulistorum*.

Paper II

In Paper II, we examine *Wolbachia* between three semispecies of *D. paulistorum*, Orinocan (OR), Amazonian (AM) and Andean-Brazilian (AB). Given the impact of *w*Pau in the prezygotic and postzygotic isolation of the semispecies, we are interested in the history of the infection in the species, i.e., whether the ancestor of all *D. paulistorum* semispecies were infected by the ancestor of all *w*Pau strains, or whether the semispecies were infected after their divergence. This would be reflected by congruency between *Wolbachia* and semispecies phylogenies. Our results show that *Wolbachia* strains from the AM semispecies form one clade, while *w*PauMS, the strain from the AB semispecies, clusters within the *Wolbachia* strains from OR, generally mirroring the phylogeny of the semispecies. The mitochondrial phylogeny is also congruent, except for the MS line, which has the α mitotype, while *w*PauMS is more related to strains co-existing with the β mitotype. To evaluate temporal congruence between *w*Pau and *D. paulistorum* semispecies diversification, we create a chronogram for the *w*Pau strains, calibrated based on the estimated time *w*Au was horizontally transferred into *D. simulans*. The divergence times for *Wolbachia* match those for the OR-AB-AM semispecies from Zanini et al., 2018 [357], and only slightly those from Paper III. *Wolbachia* mutation rates are also about one order of magnitude larger than the estimates in Paper I. Next, synteny and content differences between the *w*PauOR and *w*PauAM strains show many rearrangements around the area of the Undecim cluster, while the cluster itself is triplicated instead of duplicated, in a few *w*PauOR genomes. We identify a handful of uniquely present or absent genes among the *w*Pau strains, as well as genes with non-synonymous SNPs between *w*PauAM and *w*PauOR/AB, one of which is predicted to be a type IV secretion system effector protein. Thus, only a few differences are identified between the strains infecting the semispecies of *D. paulistorum*, that may contribute to their reproductive isolation.

Paper III

In Paper III, we focus on the relationship of N-mt and mt proteins in *D. paulistorum*. Two diverged mitotypes can be found in *D. paulistorum*, the α and the β, and both have been speculated to have been introgressed [38]. Under the compensatory model of co-evolution between N-mt and mt components, introgressed mt components are expected to be incompatible with the native N-mt proteins and reduce fitness. Thus, for introgressed mitochondria to be maintained, N-mt proteins need to remain compatible with the interacting mitochondrial proteins, either by nuclear co-introgression of compatible proteins or by fixation of *de novo* mutations in the incompatible N-mt genes. To see if this is the case, we first attempt to identify which of the mitotypes are more

likely to have been introgressed in the past, by building chronograms of the willistoni and saltans groups for their mitochondrial and nuclear genomes. Comparison of the divergence time estimates suggests that the β mitotype introgressed into *D. paulistorum*, while the α mitotype is likely the original mitotype of the species. Assuming the α mitotype being the native mitochondria for *D. paulistorum*, we looked for evidence of compensation in N-mt genes in the lines carrying the β mitotype. We evaluated two ways in which this could happen in the β mitotype lines, by co-introgression of N-mt genes or by *de novo* mutations in the nuclear background. There is scarce evidence for co-introgression between *D. paulistorum* and other species of the willistoni subgroup. Moreover, we looked for non-synonymous substitutions in N-mt proteins being fixed between α and β mitotype lines, first with more detail in OXPHOS and aaRS proteins, and then more broadly, by searching for proteins with fixed non-synonymous SNPs in N-mt proteins, between lines carrying the α and β mitotypes, and by performing selection tests for all single-copy N-mt we could recover in our protein clustering. By reconstructing ancestral sequences for some mt and N-mt coding sequences and by examining the 3D structure of proteins, we identify a possible case of compensation in the ND39 subunit of the NADH dehydrogenase of OXPHOS.

Paper IV

In Paper IV, we focus on the describing Mega-NUMTs in genomes of *D. paulistorum*. First, we search for NUMT insertions across the *D. paulistorum* genomes assembled in Paper III by using BLAST, and we create manually curated annotations for each NUMT. Mega-NUMTs are identified in six lines of *D. paulistorum* from the AB, AM, and CA semispecies. NUMTs that form the Mega-NUMTs are in tandem just like in the O11 line [215], are longer than 10kb and frequently spanning the whole mitochondrial genome, and are closely related to the α mitotype of *D. paulistorum*. The Mega-NUMTs span up to 700 copies of the mitochondrial genome. Based on their position on the genome and on synteny comparisons, we show that the Mega-NUMT in the OR and AB semispecies have their position shared, and that Mega-NUMTs between OR-AB, AM and CA have different positions. Phylogenetic analyses of the NUMTs show that they have a common origin and they diversify within each line, suggesting the insertion occurred before the divergence of the semispecies. Finally, by examining the relative genomic position of closely related NUMTs within each Mega-NUMT, we show that tandem duplication explains, at least partly, the large size of these genomic elements.

Conclusions and perspectives

The work compiled in this thesis contributes to several fields of biology. First, to genomics, by providing high-quality assemblies and annotations of nuclear, mitochondrial and endosymbiont genomes of several Neotropical fruit fly species. Second, to endosymbiosis research, by evaluating the degree of Neotropical fruit fly dependence on *Wolbachia* through testing whether host and endosymbiont evolutionary histories are linked by co-speciation or, alternatively, whether infections result from horizontal transmission. Third, to genome evolution, for which we study genomic structural changes associated with shifts in *Wolbachia* lifestyle, and the recently discovered Mega-NUMT insertions. Fourth, to broader topics of evolution, such as past introgression events between species, or the co-evolution of mitochondrial components. While not every question addressed in this thesis has been conclusively answered, I believe our work points to clear directions for future research in Neotropical *Drosophila*.

In my opinion, future research in Neotropical fruit flies should address the following. First, there is great diversity in Neotropical species for which we have little to no genomic data. For some species like *D. insularis*, there is just one genome available, and other species from more rare subgroups within the willistoni group, like the alagitans, have not been sequenced at all. Moreover, sequencing data from several populations could help us understand, whether there is strong population structure, estimate recent and ancestral effective population sizes, and infer recombination maps, all of which would be useful for interpreting the evolutionary history of the Neotropical *Drosophila*.

Second, it would be ideal to also have the genomes of *Wolbachia* from the remaining unsampled fruit flies of the willistoni group, like the CA, IN and TR semispecies of *D. paulistorum*, or from *D. insularis* and *D. nebulosa*. The phylogenetic placement of those and the resulting topology would either support or disprove co-speciation between Neotropical fruit flies and *Wolbachia* or more specifically, *D. paulistorum* semispecies and their *w*Pau strains. More sampling and sequencing may also reveal the origin of the β mitotype in *D. paulistorum* and contribute to our understanding of the past introgression events within the group.

Third, more types of data could be collected for fruit fly lines for which we have good assemblies and annotations. RNA sequencing from all semispecies

could point out semispecies-specific differences in expression relating to speciation or to different *Wolbachia* titers in the body. Hi-C data could help scaffold our assemblies more accurately and provide information on chromatin organization between lines with and without Mega-NUMTs.

Fourth would be to validate the status of semispecies in newly collecting lines. Most of the new lines of *D. paulistorum* in this work are placed into semispecies based on the nuclear molecular data, but incompatibility between them, and thus a distinct semispecies status, can only be verified after the crosses between them are done. Thus, crossing experiments between *D. paulistorum* lines would benefit studies looking into speciation of *D. paulistorum*. Moreover, it would be interesting to evaluate the relative fitness of the α and β *D. paulistorum* mitotypes conditioned on the nuclear background. This would be a great effort though as it would entail either a lot of line crosses and back-crossing between semispecies or cytoplasm transfers in embryos, which are not clear that would work for *D. paulistorum*, while also accounting for *Wolbachia* effects on host fitness.

Svensk sammanfattning

Framsteg inom DNA-sekvenseringstekniken har gjort det möjligt för forskare att undersöka organismers evolutionära historia på ett mer detaljerat sätt än någonsin tidigare. Med hjälp av bioinformatiska metoder kan genom rekonstrueras och genetiska skillnader mellan dem lokaliseras. Dessa kan användas för att identifiera tidigare hybridisering mellan arter, dra slutsatser om historiska demografiska förändringar och förstå hur naturligt urval har format dagens arter. Genomik kan också hjälpa oss att förstå den evolutionära historien hos mer komplicerade system, såsom symbios mellan två arter. Denna avhandling använder bioinformatik och genomik för att förstå den evolutionära historien hos tre olika fenomen. För det första det symbiotiska förhållandet mellan neotropiska fruktflugor och deras bakteriella endosymbionter av släktet *Wolbachia*, för det andra, tidigare hybridisering mellan dessa arter av neotropiska fruktflugor och för det tredje fruktflugornas genomutveckling i förhållande till deras mitokondrier.

Symbios mellan insekter och bakterier är vanligt i naturen och förekommer i många olika former. Relationerna kan variera från mutualistiska, där symbionterna är livsviktiga för sina värdars överlevnad, till parasitistiska, där värdar utnyttjas för symbiontens reproduktion. Bland de mer intima symbiotiska relationerna som kan utvecklas är intracellulär endosymbios, där symbionten lever inuti sin värds celler och förs vidare till nästa generation från honan till hennes avkomma. Om dessa relationer är till fördel för värden kan de vara stabila under lång tid och uppvisa tecken på samartbildning, dvs. att värdens och symbiontens artbildning är sammanlänkade. Om symbionten däremot utnyttjar värden är dessa relationer mer dynamiska och symbionten kan gå förlorad hos vissa värdarter och samtidigt spridas till andra. *Wolbachia* är en bakteriell endosymbiont som infekterar många insekter och som är känd för reproduktiv manipulation av sina värdar. Hos neotropiska *Drosophila* har reproduktionsmanipulation inte påvisats, dock har *Wolbachia* stammen *w*Pau hos arten *Drosophila paulistorum* visat sig vara obligat för värdens överlevnad. I den första och andra artikeln i denna avhandling undersöker vi därför de genomiska skillnaderna mellan den obligata *w*Pau-stammen och besläktade stammar. Vi frågar om de observerade infektionerna härrör från en lång och stabil infektion hos den gemensamma förfadern till de neotropiska

fruktflugorna, eller om de är resultatet av flera nyligen förvärvade infektioner hos de arter som observeras idag. Beträffande de genomiska skillnaderna mellan *w*Pau och andra neotropiska *Wolbachia*, finner vi att *w*Pau har fler pseudogener, fler insertioner av transposabla element och fler genomiska omarrangemang. Profagen WO som är integrerad i neotropiska *Wolbachia* och som bär gener som är ansvariga för värdmanipulation saknar dessutom många av sina strukturella gener, medan en uppsättning av elva gener, kallad Undecim-ön, är duplicerad och ibland tredubblad i *w*Pau. Den exakta funktionen hos generna i Undecim-ön är okänd, men vi spekulerar i att de spelar en viktig roll för den obligata symbiosen mellan *w*Pau och dess värdar.

Hybridisering mellan olika arter är vanligt i naturen. När hybrider återkorsas med en av föräldraarterna kan mitokondrier från en art överföras till den andra, i en process som kallas mitokondriell introgression. Mitokondrier bildar proteinkomplex som är avgörande för korrekt mitokondriell funktion och energiproduktion i cellerna. Dessa proteinkomplex består av proteiner som kodas av både kärngenomet och mitokondriegenomet. En konsekvens av mitokondriell introgression skulle således kunna vara att de mitokondriella proteinerna som överförts via introgression inte interagerar korrekt med proteiner från den nukleära bakgrunden, vilket kan leda till förlust av fitness. I sådana fall skulle mitokondriell introgression selekteras bort och inte observeras i naturen. I de fall där detta observeras kan således en av två saker ha inträffat. Antingen har kompatibla nukleära gener överförts tillsammans med mitokondrierna till den mottagande arten, eller så har de novo-mutationer i de inkompatibla nukleära proteinerna uppstått och kompenserat för de negativa fitnesseffekterna. I den tredje artikeln i denna avhandling undersöker vi introgression inom willistoni-gruppen av fruktflugor och huruvida mitokondriell introgression i *Drosophila paulistorum* har resulterat i antingen överföring av kompatibla nukleära gener eller de novo-kompensation. Vi finner bevis för överföring av kompatibla nukleära gener mellan andra neotropiska fruktflugor, men inte någon överföring som involverar *D. paulistorum*. Vi undersöker även proteiner relaterade till oxidativ fosforylering och mitokondriell translation men finner begränsat stöd för omfattande kompensation via de novo-mutationer. Vi identifierar dock ett möjligt fall av kompensation via de novo-mutationer i NADH-dehydrogenaset för oxidativ fosforylering, där substitutioner i den nukleära ND39 och den mitokondriella ND3 kan påverka bindningen av en kardiolipinligandmolekyl i proteinkomplexet.

Som nämnts har mitokondrier sitt eget genom, vilket är lokaliserat till mitokondriematrixen. Det nukleära genomet, som är begränsat till kärnan, är således fysiskt separerat från mitokondrierna, vilket hämmar all direkt interaktion mellan de två. Dessutom visar genetiska experiment i jäst att det finns

molekylära mekanismer som skyddar mot överföring av mitokondriellt DNA till kärnan, ett fenomen som kallas för mitokondrieflykt. I åldrande celler och vid sjukdomar som cancer sker mitokondrieflykt till kärnan ofta, vilket leder till att mitokondrie-DNA integreras i kärngenomet och bildar så kallade NUMT. Mekanismerna bakom bildandet av NUMT är inte väl förstådd. Nyligen har en särskild typ av NUMT-insertioner föreslagits som en förklaring till molekylära tester som tyder på faderlig överföring av mitokondrier hos människor. Den föreslagna sammanslagningen av flera mitokondriella genom integrerade i kärngenomet kallas Mega-NUMT och har nyligen bevisats förekomma i en linje av *D. paulistorum*. I den fjärde artikeln i denna avhandling använder vi genomik för att identifiera Mega-NUMT i tre underarter av *D. paulistorum*. Insertionerna kan omfatta flera hundra kopior av mitokondriegenomet och verkar ha uppstått genom att mitokondriegenomet integrerades i kärngenomet för miljontals år sedan. Vidare beskriver vi strukturen och positionen för dessa element och spekulerar kring varför detta genomiska element finns kvar i genomet.

Ελληνική περίληψη

Η συνεχής πρόοδος των τεχνολογιών αλληλούχισης DNA έχει επιτρέψει στους ερευνητές να διερευνήσουν το μακρινό παρελθόν των οργανισμών με μεγαλύτερη λεπτομέρεια από ποτέ. Ανακατασκευάζοντας γονιδιώματα χρησιμοποιώντας βιοπληροφορικές μεθόδους, μπορεί κανείς να εντοπίσει γενετικές διαφορές μεταξύ ειδών και να συμπεράνει ιστορικά γεγονότα σχετικά με τη δημογραφία τους, να εντοπίσει αρχαία συμβάντα υβριδισμού μεταξύ ειδών και να κατανοήσει πώς η φυσική επιλογή έχει διαμορφώσει τα είδη του σήμερα. Η γονιδιωματική μπορεί επίσης να μας βοηθήσει να κατανοήσουμε την εξελικτική ιστορία πιο περίπλοκων συστημάτων, όπως για παράδειγμα τη συμβίωση μεταξύ δύο ειδών. Αυτή η διατριβή χρησιμοποιεί τη βιοπληροφορική και τη γονιδιωματική για να κατανοήσει την εξελικτική ιστορία τριών πραγμάτων. Πρώτον, τη συμβιωτική σχέση μεταξύ των νεοτροπικών φρουτόμυγων και των βακτηριακών ενδοσυμβιωτών *Wolbachia*, δεύτερον, γεγονότα υβριδισμού μεταξύ των διαφορετικών ειδών νεοτροπικών φρουτόμυγων και τρίτον, την εξελικτική ιστορία του γονιδιώματος των φρουτόμυγων και των μιτοχονδρίων τους.

Η συμβίωση μεταξύ εντόμων και βακτηρίων είναι αρκετά συχνή στη φύση και εμφανίζεται σε πολλές μορφές. Μπορεί να κυμαίνεται από αμοιβαία, όπου οι βακτηριακοί ενδοσυμβιώτες είναι ζωτικής σημασίας για την επιβίωση των ξενιστών τους, έως παρασιτική, όπου οι ενδοσυμβιώτες εκμεταλλεύονται τους ξενιστές τους για αναπαραγωγικό όφελός τους. Η πιο στενή από τις συμβιωτικές σχέσεις που μπορούν να εξελιχθούν είναι αυτή της ενδοκυτταρικής ενδοσυμβίωσης, όπου ο ενδοσυμβιώτης κατοικεί μέσα στα κύτταρα των ξενιστών του και κληρονομείται στην επόμενη γενιά ξενιστών από τη μητέρα στους απογόνους. Εάν αυτές οι σχέσεις παρέχουν κάποιο όφελος για τους ξενιστές τους, μπορούν να είναι σταθερές για μεγάλο χρονικό διάστημα και να παρουσιάζουν σημάδια συν-ειδογένεσης, όπου η ειδογένεση ξενιστών και συμβιωτών συνδέονται μεταξύ τους και συμβαίνουν σχεδόν ταυτόχρονα. Εάν οι συμβιωτικές σχέσεις είναι εκμεταλλευτικές προς τον ξενιστή, τότε εμφανίζονται ως πιο δυναμικές, με τον ενδοσυμβιώτη να χάνεται από ορισμένα είδη ξενιστών και να μολύνει άλλα νέα είδη. Η *Wolbachia* είναι βακτηριακός ενδοσυμβιώτης πολλών εντόμων και γνωστή για τον αναπαραγωγικό χειρισμό των ξενιστών της. Ωστόσο, στις νεοτροπικές

φρουτόμυγες με *Wolbachia* δεν έχει αποδειχθεί αναπαραγωγική χειραγώγηση, και σε ένα είδος αυτών των φρουτόμυγων, το *D. paulistorum*, το *w*Pau στέλεχος της *Wolbachia* έχει αποδειχθεί ότι είναι υποχρεωτικό για την επιβίωση του ξενιστή του. Έτσι, στην πρώτη και δεύτερη εργασία αυτής της διατριβής εξετάζουμε τις γονιδιωματικές διαφορές μεταξύ του υποχρεωτικού στελέχους *w*Pau έναντι άλλων συγγενών στελεχών και διερωτόμαστε εάν οι μολύνσεις είναι αποτέλεσμα μιας μακράς και σταθερής προγονικής μόλυνσης στον πρόγονο των νεοτροπικών φρουτόμυγων ή εάν είναι αποτέλεσμα πολλαπλών πρόσφατων συμβάντων μόλυνσης στο είδος που παρατηρούνται σήμερα.

Ο υβριδισμός μεταξύ διαφορετικών ειδών είναι συνήθης στη φύση. Όταν υβρίδια διασταυρώνονται με κάποιο από τα γονικά είδη, τα μιτοχόνδρια από το ένα γονικό είδος μπορούν να μεταφερθούν στο άλλο, με τη διαδικασία αυτή να ονομάζεται μιτοχονδριακή εισβολή. Τα μιτοχόνδρια παράγουν πρωτεϊνικά σύμπλοκα που είναι ζωτικής σημασίας για τη σωστή μιτοχονδριακή λειτουργία και την παραγωγή ενέργειας των κυττάρων. Αυτά τα πρωτεϊνικά σύμπλοκα αποτελούνται από πρωτεΐνες που κωδικοποιούνται τόσο από τον πυρήνα όσο και από τα μιτοχονδριακή γονιδιώματα. Έτσι, μια συνέπεια της μιτοχονδριακής εισβολής θα μπορούσε να είναι ότι οι εισβάλουσες μιτοχονδριακές πρωτεΐνες δεν σχηματίζουν σωστές αλληλεπιδράσεις με πρωτεΐνες του πυρηνικού υποβάθρου, προκαλώντας τη δυσλειτουργία των πρωτεϊνικών συμπλόκων. Σε αυτήν την περίπτωση, η μιτοχονδριακή διείσδυση δεν θα ευνοούνταν από τη φυσική επιλογή και δεν θα παρατηρούνταν στη φύση. Για τις περιπτώσεις όμως που παρατηρείται, οι πιθανές εξηγήσεις είναι δύο. Είτε συμβατά πυρηνικά γονίδια να συν-εισβάλουν με τα μιτοχόνδρια στο είδος-αποδέκτη, είτε *de novo* μεταλλάξεις στις ασύμβατες πυρηνικές πρωτεΐνες να πραγματοποιηθούν και να αντισταθμίσουν τις αρνητικές επιδράσεις της μιτοχονδριακής εισβολής. Έτσι λοιπόν, στην τρίτη εργασία αυτής της διατριβής, εξετάζουμε συμβάντα γενετικής εισβολής εντός της ομάδας willistoni των φρουτόμυγων και εάν μια μιτοχονδριακή εισβολή στη *Drosophila paulistorum* έχει οδηγήσει είτε σε πυρηνική συν-εισαγωγή είτε σε *de novo* αντιστάθμιση. Όσον αφορά τις γονιδιωματικές διαφορές μεταξύ του *w*Pau και άλλων νεοτροπικών *Wolbachia*, διαπιστώνουμε ότι το γονιδίωμα του *w*Pau έχει περισσότερα ψευδογονίδια, περισσότερες ενθέσεις μεταθετών στοιχείων και περισσότερες γονιδιωματικές αναδιατάξεις. Επιπλέον, ο προφάγος WO που έχει εισαχθεί στη νεοτροπική *Wolbachia* και φέρει γονίδια υπεύθυνα για τον χειρισμό του ξενιστή, έχει τα περισσότερα του δομικά γονίδια απόντα, αλλά έχει παρών ένα σύνολο έντεκα γονιδίων, που ονομάζεται νησίδα Undecim, το οποίο έχει διπλασιαστεί και μερικές φορές τριπλασιαστεί στο *w*Pau. Η ακριβής

λειτουργία των γονιδίων στη νησίδα Undecim είναι άγνωστη, αλλά εικάζουμε ότι παίζουν σημαντικό ρόλο στην υποχρεωτική συμβίωση του *w*Pau για τους ξενιστές του.

Όπως προαναφέρθηκε, τα μιτοχόνδρια έχουν το δικό τους γονιδίωμα, το οποίο περιορίζεται στην μιτοχονδριακή θεμέλια ουσία. Από την άλλη, το πυρηνικό γονιδίωμα περιορίζεται στον πυρήνα και είναι χωρικά διαχωρισμένο από το μιτοχονδριακό, παρεμποδίζοντας έτσι οποιαδήποτε άμεση αλληλεπίδραση μεταξύ των δύο. Επιπλέον, γενετικά πειράματα σε ζυμομύκητες δείχνουν ότι υπάρχουν μοριακοί μηχανισμοί που είναι υπεύθυνοι για την προστασία του πυρήνα από τη εισαγωγή μιτοχονδριακού DNA, μια διαδικασία που ονομάζεται μιτοχονδριακή διαφυγή. Σε γηράσκοντα κύτταρα και σε ασθένειες όπως ο καρκίνος, η μιτοχονδριακή διαφυγή στον πυρήνα συμβαίνει συχνά, δημιουργώντας πυρηνικές-μιτοχονδριακές ενθέσεις ή NUMTs. Ο μηχανισμός με τον οποίο δημιουργούνται οι NUMTs δεν είναι καλά κατανοητός. Πρόσφατα, έχει προταθεί ένας ειδικός τύπος NUMT για να εξηγήσει αποτελέσματα μοριακών εξετάσεων που υποδεικνύουν πατρική μετάδοση μιτοχονδρίων στους ανθρώπους. Αυτή η πρόταση περιγράφει τη συνένωση πολλαπλών μιτοχονδριακών γονιδιωμάτων και την εισαγωγή τους στο πυρηνικό γονιδίωμα. Αυτό το συνονθύλευμα μιτοχονδριακών γονιδιωμάτων ονομάζεται Mega-NUMT και η ύπαρξή του αποδείχθηκε πρόσφατα σε μία εργαστηριακή σειρά της *D. paulistorum*. Στην τέταρτη εργασία αυτής της διατριβής, χρησιμοποιούμε γονιδιωματικές μεθόδους για να εντοπίσουμε Mega-NUMTs σε τρία ημιείδη της *D. paulistorum*. Οι μιτοχονδριακές ενθέσεις μπορεί να έχουν έκταση αρκετών εκατοντάδων αντιγράφων του μιτοχονδριακού γονιδιώματος και φαίνεται να είναι το αποτέλεσμα της εισαγωγής ενός μιτοχονδριακού γονιδιώματος πριν από εκατομμύρια χρόνια. Περιγράφουμε περαιτέρω τη δομή και τη θέση αυτών των στοιχείων και εικάζουμε τους λόγους για τους οποίους το Mega-NUMT παραμένει στο γονιδίωμα.

Acknowledgements

My PhD studies have been quite the journey. I believe I have learned a lot along the way, and I am very happy to have reached this far. At this point I would like to thank the people, without which, this would have been impossible.

First Lisa. I have been saying it for years now to anyone asking me about my studies, that is, one of the best things about this PhD is that I got the best supervisor that I could possibly get. You have been a great mentor to me, by always being available to discuss anything, by encouraging me to learn and try new things, by handling my stubbornness and presenting to me the alternative of cooperativeness, by supporting me at times I felt I should give up, by reminding me that I have being doing well along the way, and by setting the standard for me to reach, should the day come and I need to mentor someone myself. Thank you.

Οι επόμενοι δύο χωρίς τους οποίους δεν θα κατάφερνα να φτάσω ως εδώ είναι ο μπαμπάς και ο θείος μου. Με στηρίξατε στο ξεκίνημά μου για σπουδές σε μια χώρα πολύ μακρινή και διαφορετική χώρα από την Ελλάδα, με αποτέλεσμα να ζήσω μια ενήλικη ζωή στο εξωτερικό, να γνωρίσω διαφορετικές κουλτούρες ανθρώπων απ' όλη την υφήλιο και να καταφέρω να φτάσω ένα επίπεδο κατάρτισης στη δουλεία μου που δεν θα το έφτανα ποτέ στην Ελλάδα. Παρόλο που δεν ξέρω ακόμα που θέλω να βρίσκομαι και τι θα κάνω στο μέλλον, όλα αυτά μου έχουν δώσει μια προοπτική των πραγμάτων που σίγουρα θα με βοηθήσει να επιλέξω κατεύθυνση για το ύστερα, κάτι που δεν θα είχα αν δεν ήσαστε εσείς. Σας αγαπώ, μου έχετε λείψει, και ανυπομονώ να επιστρέψω να σας δω.

Here, I would like to thank people who have contributed to my studies. First, my co-supervisor Lionel, who has been supportive throughout the years and helped me wrap my head around chronograms. Second, Guilherme who has been providing great advice and feedback on how to perform genome annotations in *Drosophila*. Third, Linnéa for doing an amazing job on the third paper for this thesis. I wish you the best in your future defence! I would also like to thank Linnéa, Burçin, Lorenzo and Humberto for reading my thesis draft and providing valuable feedback.

Finally, I would like to mention my friends that were there for me during both the fun and the frustrating moments.

Lorenzo, Patrick and Filip, I was very fortunate to have met you and spent the past years together with you. I hope that the post-PhD life that you have started brings you the best and that we will meet again all together soon. Miss you.

Στη Βασούλα και τη Σοφία που ξεκινήσαμε μαζί από Αθήνα για Σουηδία και θα έχουν το defence τους πολύ σύντομα, είμαι σίγουρος ότι όλα θα πάνε τέλεια, γιατί είστε αφοσιωμένες στον στόχο σας και αγαπάτε αυτό που κάνετε (όσο αγαπιέται το διδακτορικό τέλος πάντων). Σας εύχομαι λοιπόν σύντομα να ξεμπερδεύετε.

Στην Αγάπη που είναι πρώην μαθήτριά μου, καλή μου φίλη, και που ξεκίνησε πρόσφατα το διδακτορικό της, σου εύχομαι πολλές επιτυχίες για τα χρόνια που έρχονται και ελπίζω να βρεθούμε σύντομα από κοντά, είτε στην Ισπανία είτε στην Κρήτη.

To the gym bro, Hugo, thank you for the evenings we spent together breaking records, we must keep it up. We cannot be defeated!

Humberto, Nik and Violeta, despite having spent only a limited time with you, I enjoyed it very much and I wish we could have met earlier. For the latter two, good luck with your PhDs, I'm sure you will do great things. Humberto, thank you for all the fun times with the board games and for introducing me to the best cat in the world, Mr. Felix!

Στους ετέρους βιολόγους, Αλέξανδρο, Σπύρο, Σοφοκλή και Σοφία, ο καθένας σας είναι σε διαφορετικό στάδιο με τις σπουδές του, αλλά σας εύχομαι αντιστοίχως, να ξεκινήσουν, να ολοκληρωθούν και να οδηγήσουν σε επιτυχίες στο μέλλον. Ανυπομονώ να σας ξαναδώ.

Στους δύο καθυστερημένους που έχω για κολλητούς, τον Αλέξανδρο και τον Περικλή, παρόλο που βρισκόμαστε μακριά ο ένας από τον άλλον όλα αυτά τα χρόνια, είστε εκεί και με στηρίζετε είτε διαδικτυακά είτε από κοντά όταν βρισκόμαστε στις διακοπές μας. Χαίρομαι πολύ που βλέπω τα πράγματα στη ζωή σας να μπαίνουν σιγά σιγά σε μια σειρά και ελπίζω να τα φέρει έτσι η ζωή κάποια στιγμή ώστε να είμαστε γείτονες ξανά σαν 15 χρόνια πριν. Σας έχω στην καρδιά μου.

References

1. O'Grady PM, DeSalle R. Phylogeny of the Genus Drosophila. Genetics. 2018;209:1–25. https://doi.org/10.1534/genetics.117.300583.
2. Markow TA, O'Grady P. Drosophila: A Guide to Species Identification and Use. Chantilly, UNITED STATES: Elsevier Science & Technology; 2005.
3. Kim BY, Gellert HR, Church SH, Suvorov A, Anderson SS, Barmina O, et al. Single-fly genome assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life. PLOS Biol. 2024;22:e3002697. https://doi.org/10.1371/journal.pbio.3002697.
4. Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, et al. Highly contiguous assemblies of 101 drosophilid genomes. eLife. 2021;10:e66405. https://doi.org/10.7554/eLife.66405.
5. Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, et al. Widespread introgression across a phylogeny of 155 *Drosophila* genomes. Curr Biol. 2022;32:111-123.e5. https://doi.org/10.1016/j.cub.2021.10.052.
6. Carson HL. The Ecology of Drosophila breeding sites. Harold Lyon Arbor Lect. 1971;2:1–28. https://doi.org/10.5281/zenodo.10738509.
7. Hardin JA, Kraus DA, Burrack HJ. Diet quality mitigates intraspecific larval competition in Drosophila suzukii. Entomol Exp Appl. 2015;156:59–65. https://doi.org/10.1111/eea.12311.
8. Lewis MT, Hamby KA. Differential Impacts of Yeasts on Feeding Behavior and Development in Larval Drosophila suzukii (Diptera:Drosophilidae). Sci Rep. 2019;9:13370. https://doi.org/10.1038/s41598-019-48863-1.
9. Anagnostou C, Dorsch M, Rohlfs M. Influence of dietary yeasts on Drosophila melanogaster life-history traits. Entomol Exp Appl. 2010;136:1–11. https://doi.org/10.1111/j.1570-7458.2010.00997.x.
10. Leitão-Gonçalves R, Carvalho-Santos Z, Francisco AP, Fioreze GT, Anjos M, Baltazar C, et al. Commensal bacteria and essential amino acids control food choice behavior and reproduction. PLOS Biol. 2017;15:e2000862. https://doi.org/10.1371/journal.pbio.2000862.
11. Markow TA, O'Grady P. Reproductive Ecology of Drosophila. Funct Ecol. 2008;22:747–59.
12. O'Grady PM, Kidwell MG. Phylogeny of the Subgenus Sophophora (Diptera: Drosophilidae) Based on Combined Analysis of Nuclear and Mitochondrial Sequences. Mol Phylogenet Evol. 2002;22:442–53. https://doi.org/10.1006/mpev.2001.1053.
13. Bächli G. TaxoDros: The Database on Taxonomy of Drosophilidae. TaxoDros. 2026. https://doi.org/10.5281/zenodo.18764503.
14. Zanini R, Deprá M, Valante V. On the geographic distribution of the Drosophila willistoni group (Diptera, Drosophilidae) ? updated geographic distribution of the Neotropical willistoni subgroup. Drosoph Inf Serv. 2015;98:39–43.

15. Zanini R, Deprá M, Valente V. On the geographic distribution of the Drosophila willistoni group (Diptera, Drosophilidae) - updated distribution of alagitans and bocainensis subgroups. Drosoph Inf Serv. 2015;98:25–7.
16. Kastritsis CD, Dobzhansky T. Drosophila pavlovskiana, a Race or a Species? Am Midl Nat. 1967;78:244–8. https://doi.org/10.2307/2423385.
17. Dobzhansky T, Ehrman L, Pavlovsky O. Drosophila insularis, a New Sibling Species of the willistoni Group. Univ Tex Publ. 1957;5721:39–47.
18. Townsend JI. Cryptic Subspeciation in Drosophila Belonging to the Subgenus Sophophora. Am Nat. 1954;88:339–51. https://doi.org/10.1086/281842.
19. Ayala FJ, Tracey ML, Barr LG, Ehrenfeld JG. GENETIC AND REPRODUCTIVE DIFFERENTIATION OF THE SUBSPECIES, DROSOPHILA EQUINOXIALIS CARIBBENSIS. Evolution. 1974;28:24–41. https://doi.org/10.1111/j.1558-5646.1974.tb00723.x.
20. Mardiros XB, Park R, Clifton B, Grewal G, Khizar AK, Markow TA, et al. Post-mating Reproductive isolation between strains of Drosophila willistoni. Fly (Austin). 2016;10:162–71. https://doi.org/10.1080/19336934.2016.1197448.
21. Dobzhansky T, Spassky B. Drosophila Paulistorum, a Cluster of Species in Statu Nascendi. Proc Natl Acad Sci. 1959;45:419–28. https://doi.org/10.1073/pnas.45.3.419.
22. Perez-Salas S, Richmond RC, Pavlovsky O, Kastritsis CD, Ehrman L, Dobzhansky Th. The Interior Semispecies of Drosophila paulistorum. Evolution. 1970;24:519–27. https://doi.org/10.2307/2406831.
23. Ehrman L. Direct Observation of Sexual Isolation between Allopatric and between Sympatric Strains of the Different Drosophila paulistorum Races. Evolution. 1965;19:459–64. https://doi.org/10.2307/2406243.
24. O'Hara E, Pruzan A, Ehrman L. Ethological isolation and mating experience in Drosophila paulistorum. Proc Natl Acad Sci. 1976;73:975–6. https://doi.org/10.1073/pnas.73.3.975.
25. Kim Y-K, Ehrman L, Koepfer HR. Developmental isolation and subsequent adult behavior ofDrosophila paulistorum. I. Survey of the six semispecies. Behav Genet. 1992;22:545–56. https://doi.org/10.1007/BF01074307.
26. Dobzhansky Th, Ehrman L, Pavlovsky O, Spassky B. THE SUPERSPECIES DROSOPHILA PAULISTORUM. Proc Natl Acad Sci. 1964;51:3–9. https://doi.org/10.1073/pnas.51.1.3.
27. Kim Y-K, Phillips DR, Chao T, Ehrman L. Developmental Isolation and Subsequent Adult Behavior of Drosophila paulistorum. VI. Quantitative Variation in Cuticular Hydrocarbons. Behav Genet. 2004;34:385–94. https://doi.org/10.1023/B:BEGE.0000023644.87050.1a.
28. Schneider DI, Ehrman L, Engl T, Kaltenpoth M, Hua-Van A, Le Rouzic A, et al. Symbiont-Driven Male Mating Success in the Neotropical Drosophila paulistorum Superspecies. Behav Genet. 2019;49:83–98. https://doi.org/10.1007/s10519-018-9937-8.
29. Miller WJ, Ehrman L, Schneider D. Infectious Speciation Revisited: Impact of Symbiont-Depletion on Female Fitness and Mating Behavior of Drosophila paulistorum. PLoS Pathog. 2010;6:e1001214. https://doi.org/10.1371/journal.ppat.1001214.
30. Baião GC, Schneider DI, Miller WJ, Klasson L. Multiple introgressions shape mitochondrial evolutionary history in Drosophila paulistorum and the Drosophila willistoni group. Mol Phylogenet Evol. 2023;180:107683. https://doi.org/10.1016/j.ympev.2022.107683.

31. Valadão H, Proença CEB, Kuhlmann MP, Harris SA, Tidon R. Fruit-breeding drosophilids (Diptera) in the Neotropics: playing the field and specialising in generalism? Ecol Entomol. 2019;44:721–37. https://doi.org/10.1111/een.12769.
32. Camila Heloise dos Santos, Karoline Aparecida Vieira dos Santos, Luciana Paes de Barros Machado, Rogério Pincela Mateus. Life History Traits and Metabolic Pool Variation in Neotropical Species of Drosophila (Diptera, Drosophilidae). Zool Stud. 2023;無. https://doi.org/10.6620/ZS.2023.62-56.
33. Ma S, Avanesov AS, Porter E, Lee BC, Mariotti M, Zemskaya N, et al. Comparative transcriptomics across 14 Drosophila species reveals signatures of longevity. Aging Cell. 2018;17:e12740. https://doi.org/10.1111/acel.12740.
34. Machida W, Tidon R, Klaczko J. Wing plastic response to temperature variation in two distantly related Neotropical Drosophila species (Diptera, Drosophilidae). Can J Zool. 2021;100. https://doi.org/10.1139/cjz-2021-0099.
35. Alves Da Mata R, Tidon R. The relative roles of habitat heterogeneity and disturbance in drosophilid assemblages (Diptera, Drosophilidae) in the Cerrado. Insect Conserv Divers. 2013;6:663–70. https://doi.org/10.1111/icad.12020.
36. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007;450:203–18. https://doi.org/10.1038/nature06341.
37. Sottolano CJ, Revaitis NT, Geneva AJ, Yakoby N. Nebulous without white: annotated long-read genome assembly and CRISPR/Cas9 genome engineering in Drosophila nebulosa. G3 GenesGenomesGenetics. 2022;12:jkac231. https://doi.org/10.1093/g3journal/jkac231.
38. Baião GC, Schneider DI, Miller WJ, Klasson L. Multiple introgressions shape mitochondrial evolutionary history in Drosophila paulistorum and the Drosophila willistoni group. Mol Phylogenet Evol. 2023;180:107683. https://doi.org/10.1016/j.ympev.2022.107683.
39. Prediger C, Ferreira EA, Zorzato SV, Hua-Van A, Klasson L, Miller WJ, et al. Saltational Episodes of Reticulate Evolution in the Drosophila saltans Species Group. Mol Biol Evol. 2024;41:msae250. https://doi.org/10.1093/molbev/msae250.
40. Gebert D, Hay AD, Hoang JP, Gibbon AE, Henderson IR, Teixeira FK. Analysis of 30 chromosome-level Drosophila genome assemblies reveals dynamic evolution of centromeric satellite repeats. Genome Biol. 2025;26:63. https://doi.org/10.1186/s13059-025-03527-4.
41. Dhakad P, Kim BY, Petrov DA, Obbard DJ. Comparative gene annotation and orthology assignments across 301 species of Drosophilidae. PLOS Biol. 2026;24:e3003663. https://doi.org/10.1371/journal.pbio.3003663.
42. Oulhen N, Schulz BJ, Carrier TJ. English translation of Heinrich Anton de Bary's 1878 speech, 'Die Erscheinung der Symbiose' ('De la symbiose'). Symbiosis. 2016;69:131–9. https://doi.org/10.1007/s13199-016-0409-8.
43. Dimijian GG. Evolving together: the biology of symbiosis, part 1. Proc Bayl Univ Med Cent. 2000;13:217–26.
44. Ferrari J, Vavre F. Bacterial symbionts in insects or the story of communities affecting communities. Philos Trans R Soc B Biol Sci. 2011;366:1389–400. https://doi.org/10.1098/rstb.2010.0226.
45. Masson F, Lemaitre B. Growing Ungrowable Bacteria: Overview and Perspectives on Insect Symbiont Culturability. Microbiol Mol Biol Rev. 2020;84:10.1128/mmbr.00089-20. https://doi.org/10.1128/mmbr.00089-20.
46. Husnik F, Keeling PJ. The fate of obligate endosymbionts: reduction, integration, or extinction. Curr Opin Genet Dev. 2019;58–59:1–8. https://doi.org/10.1016/j.gde.2019.07.014.

47. Moya A, Peretó J, Gil R, Latorre A. Learning how to live together: genomic insights into prokaryote–animal symbioses. Nat Rev Genet. 2008;9:218–29. https://doi.org/10.1038/nrg2319.
48. Bright M, Bulgheresi S. A complex journey: transmission of microbial symbionts. Nat Rev Microbiol. 2010;8:218–30. https://doi.org/10.1038/nrmicro2262.
49. Hammer TJ. Why do hosts malfunction without microbes? Missing benefits versus evolutionary addiction. Trends Microbiol. 2024;32:132–41. https://doi.org/10.1016/j.tim.2023.07.012.
50. Perlmutter JI, Bordenstein SR. Microorganisms in the reproductive tissues of arthropods. Nat Rev Microbiol. 2020;18:97–111. https://doi.org/10.1038/s41579-019-0309-z.
51. Moran NA, McCutcheon JP, Nakabachi A. Genomics and Evolution of Heritable Bacterial Symbionts. Annu Rev Genet. 2008;42:165–90. https://doi.org/10.1146/annurev.genet.41.110306.130119.
52. Funk DJ, Helbling L, Wernegreen JJ, Moran NA. Intraspecific Phylogenetic Congruence among Multiple Symbiont Genomes. Proc Biol Sci. 2000;267:2517–21.
53. Zhou J, Guo Q, Han X, Zhang W, Huang Z, Dietrich CH, et al. Genome degradation results in nested symbiosis and endosymbiont replacement in cicadas. Nat Commun. 2025;16:10104. https://doi.org/10.1038/s41467-025-65129-9.
54. Baumann P. BIOLOGY OF BACTERIOCYTE-ASSOCIATED ENDOSYMBIONTS OF PLANT SAP-SUCKING INSECTS. Annu Rev Microbiol.2005;59:155–89. https://doi.org/10.1146/annurev.micro.59.030804.121041.
55. Braendle C, Miura T, Bickel R, Shingleton AW, Kambhampati S, Stern DL. Developmental Origin and Evolution of Bacteriocytes in the Aphid–Buchnera Symbiosis. PLOS Biol. 2003;1:e21. https://doi.org/10.1371/journal.pbio.0000021.
56. Noda T, Okude G, Meng X-Y, Koga R, Moriyama M, Fukatsu T. Bacteriocytes and Blattabacterium Endosymbionts of the German Cockroach Blattella germanica, the Forest Cockroach Blattella nipponica, and Other Cockroach Species. Zoolog Sci. 2020;37:399–410. https://doi.org/10.2108/zs200054.
57. Toju H, Tanabe AS, Notsu Y, Sota T, Fukatsu T. Diversification of endosymbiosis: replacements, co-speciation and promiscuity of bacteriocyte symbionts in weevils. ISME J. 2013;7:1378–90. https://doi.org/10.1038/ismej.2013.27.
58. Pais R, Lohs C, Wu Y, Wang J, Aksoy S. The Obligate Mutualist Wigglesworthia glossinidia Influences Reproduction, Digestion, and Immunity Processes of Its Host, the Tsetse Fly. Appl Environ Microbiol. 2008;74:5965–74. https://doi.org/10.1128/AEM.00741-08.
59. Perotti MA, Allen JM, Reed DL, Braig HR. Host-symbiont interactions of the primary endosymbiont of human head and body lice. FASEB J. 2007;21:1058–66. https://doi.org/10.1096/fj.06-6808com.
60. Russell JA, Latorre A, Sabater-Muñoz B, Moya A, Moran NA. Side-stepping secondary symbionts: widespread horizontal transfer across and beyond the Aphidoidea. Mol Ecol. 2003;12:1061–75. https://doi.org/10.1046/j.1365-294X.2003.01780.x.
61. Manzano-Marín A, Coeur d'Acier A, Clamens AL, Cruaud C, Barbe V, Jousselin E. Co-obligate symbioses have repeatedly evolved across aphids, but partner identity and nutritional contributions vary across lineages. Peer Community J. 2023;3:e46. https://doi.org/10.24072/pcjournal.278.
62. Corretto E, Štarhová Serbina L, Dittmer J, Michalik A, Schuler H. Evolutionary dynamics of obligate endosymbiosis in the psyllid genus Cacopsylla. Commun Biol. 2025;8:1540. https://doi.org/10.1038/s42003-025-08904-0.

63. Gottlieb Y, Ghanim M, Gueguen G, Kontsedalov S, Vavre F, Fleury F, et al. Inherited intracellular ecosystem: symbiotic bacteria share bacteriocytes in whiteflies. FASEB J. 2008;22:2591–9. https://doi.org/10.1096/fj.07-101162.
64. Meseguer AS, Manzano-Marín A, Coeur d'Acier A, Clamens A-L, Godefroid M, Jousselin E. Buchnera has changed flatmate but the repeated replacement of co-obligate symbionts is not associated with the ecological expansions of their aphid hosts. Mol Ecol. 2017;26:2363–78. https://doi.org/10.1111/mec.13910.
65. Duron O, Bouchon D, Boutin S, Bellamy L, Zhou L, Engelstädter J, et al. The diversity of reproductive parasites among arthropods: Wolbachiado not walk alone. BMC Biol. 2008;6:27. https://doi.org/10.1186/1741-7007-6-27.
66. Engelstädter J, Hurst GDD. The Ecology and Evolution of Microbes that Manipulate Host Reproduction. Annu Rev Ecol Evol Syst. 2009;40:127–49. https://doi.org/10.1146/annurev.ecolsys.110308.120206.
67. Ahmed MZ, Breinholt JW, Kawahara AY. Evidence for common horizontal transmission of Wolbachia among butterflies and moths. BMC Evol Biol. 2016;16:118. https://doi.org/10.1186/s12862-016-0660-x.
68. Binetruy F, Bailly X, Chevillon C, Martin OY, Bernasconi MV, Duron O. Phylogenetics of the *Spiroplasma ixodetis* endosymbiont reveals past transfers between ticks and other arthropods. Ticks Tick-Borne Dis. 2019;10:575–84. https://doi.org/10.1016/j.ttbdis.2019.02.001.
69. Vancaester E, Blaxter M. Phylogenomic analysis of Wolbachia genomes from the Darwin Tree of Life biodiversity genomics project. PLOS Biol. 2023;21:e3001972. https://doi.org/10.1371/journal.pbio.3001972.
70. Sanaei E, Charlat S, Engelstädter J. *Wolbachia* host shifts: routes, mechanisms, constraints and evolutionary consequences. Biol Rev. 2021;96:433–53. https://doi.org/10.1111/brv.12663.
71. Drew GC, Stevens EJ, King KC. Microbial evolution and transitions along the parasite–mutualist continuum. Nat Rev Microbiol. 2021;19:623–38. https://doi.org/10.1038/s41579-021-00550-7.
72. Whittle M, Barreaux AMG, Bonsall MB, Ponton F, English S. Insect-host control of obligate, intracellular symbiont density. Proc R Soc B Biol Sci. 2021;288:20211993. https://doi.org/10.1098/rspb.2021.1993.
73. Strunov A, Schmidt K, Kapun M, Miller WJ. Restriction of *Wolbachia* Bacteria in Early Embryogenesis of Neotropical *Drosophila* Species via Endoplasmic Reticulum-Mediated Autophagy. mBio. 2022;13:e03863-21. https://doi.org/10.1128/mbio.03863-21.
74. Wilkinson TL, Koga R, Fukatsu T. Role of Host Nutrition in Symbiont Regulation: Impact of Dietary Nitrogen on Proliferation of Obligate and Facultative Bacterial Endosymbionts of the Pea Aphid Acyrthosiphon pisum. Appl Environ Microbiol. 2007;73:1362–6. https://doi.org/10.1128/AEM.01211-06.
75. Simonet P, Duport G, Gaget K, Weiss-Gayet M, Colella S, Febvay G, et al. Direct flow cytometry measurements reveal a fine-tuning of symbiotic cell dynamics according to the host developmental needs in aphid symbiosis. Sci Rep. 2016;6:19967. https://doi.org/10.1038/srep19967.
76. Login FH, Balmand S, Vallier A, Vincent-Monégat C, Vigneron A, Weiss-Gayet M, et al. Antimicrobial Peptides Keep Insect Endosymbionts Under Control. Science. 2011;334:362–5. https://doi.org/10.1126/science.1209728.
77. Vigneron A, Masson F, Vallier A, Balmand S, Rey M, Vincent-Monégat C, et al. Insects Recycle Endosymbionts when the Benefit Is Over. Curr Biol. 2014;24:2267–73. https://doi.org/10.1016/j.cub.2014.07.065.

78. Simonet P, Gaget K, Balmand S, Ribeiro Lopes M, Parisot N, Buhler K, et al. Bacteriocyte cell death in the pea aphid/Buchnera symbiotic system. Proc Natl Acad Sci U S A. 2018;115:E1819–28. https://doi.org/10.1073/pnas.1720237115.
79. Sachs JL, Skophammer RG, Regus JU. Evolutionary transitions in bacterial symbiosis. Proc Natl Acad Sci. 2011;108 supplement_2:10800–7. https://doi.org/10.1073/pnas.1100304108.
80. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 2012;10:13–26. https://doi.org/10.1038/nrmicro2670.
81. Hrdina A, Serra Canales M, Arias-Rojas A, Frahm D, Iatsenko I. The endosymbiont Spiroplasma poulsonii increases Drosophila melanogaster resistance to pathogens by enhancing iron sequestration and melanization. mBio. 2024;15:e00936-24. https://doi.org/10.1128/mbio.00936-24.
82. Jaenike J, Unckless R, Cockburn SN, Boelio LM, Perlman SJ. Adaptation via Symbiosis: Recent Spread of a Drosophila Defensive Symbiont. Science. 2010;329:212–5. https://doi.org/10.1126/science.1188235.
83. Strunov A, Lerch S, Blanckenhorn WU, Miller WJ, Kapun M. Complex effects of environment and Wolbachia infections on the life history of Drosophila melanogaster hosts. J Evol Biol. 2022;35:788–802. https://doi.org/10.1111/jeb.14016.
84. Hedges LM, Brownlie JC, O'Neill SL, Johnson KN. *Wolbachia* and Virus Protection in Insects. Science. 2008;322:702–702. https://doi.org/10.1126/science.1162418.
85. Martinez J, Tolosana I, Ok S, Smith S, Snoeck K, Day JP, et al. Symbiont strain is the main determinant of variation in *Wolbachia* -mediated protection against viruses across *Drosophila* species. Mol Ecol. 2017;26:4072–84. https://doi.org/10.1111/mec.14164.
86. Mira A, Moran NA. Estimating Population Size and Transmission Bottlenecks in Maternally Transmitted Endosymbiotic Bacteria. Microb Ecol. 2002;44:137–43. https://doi.org/10.1007/s00248-002-0012-9.
87. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009;10:195–205. https://doi.org/10.1038/nrg2526.
88. Moran NA. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A. 1996;93:2873–8. https://doi.org/10.1073/pnas.93.7.2873.
89. Wernegreen JJ. Endosymbiont evolution: predictions from theory and surprises from genomes: Endosymbiont genome evolution. Ann N Y Acad Sci. 2015;1360:16–35. https://doi.org/10.1111/nyas.12740.
90. Muller HJ. The relation of recombination to mutational advance. Mutat Res. 1964;1:2–9. https://doi.org/10.1016/0027-5107(64)90047-8.
91. Fisher RM, Henry LM, Cornwallis CK, Kiers ET, West SA. The evolution of host-symbiont dependence. Nat Commun. 2017;8:15973. https://doi.org/10.1038/ncomms15973.
92. Renoz F, Arai H, Pons I. The genus Sodalis as a resource for understanding the multifaceted evolution of bacterial symbiosis in insects. Symbiosis. 2024;92:187–208. https://doi.org/10.1007/s13199-023-00966-0.
93. Naito M, Pawlowska TE. Defying Muller's Ratchet: Ancient Heritable Endobacteria Escape Extinction through Retention of Recombination and Genome Plasticity. mBio. 2016;7:e02057-15. https://doi.org/10.1128/mBio.02057-15.
94. Takeuchi N, Kaneko K, Koonin EV. Horizontal Gene Transfer Can Rescue Prokaryotes from Muller's Ratchet: Benefit of DNA from Dead Cells and Population Subdivision. G3 GenesGenomesGenetics. 2014;4:325–39. https://doi.org/10.1534/g3.113.009845.

95. Nakabachi A, Moran NA. Extreme Polyploidy of Carsonella, an Organelle-Like Bacterium with a Drastically Reduced Genome. Microbiol Spectr. 2022;10:e00350-22. https://doi.org/10.1128/spectrum.00350-22.
96. Komaki K, Ishikawa H. Genomic copy number of intracellular bacterial symbionts of aphids varies in response to developmental stage and morph of their host. Insect Biochem Mol Biol. 2000;30:253–8. https://doi.org/10.1016/S0965-1748(99)00125-3.
97. O'Neill SL, Giordano R, Colbert AM, Karr TL, Robertson HM. 16S rRNA phylogenetic analysis of the bacterial endosymbionts associated with cytoplasmic incompatibility in insects. Proc Natl Acad Sci. 1992;89:2699–702. https://doi.org/10.1073/pnas.89.7.2699.
98. Sironi M, Bandi C, Sacchi L, Sacco BD, Damiani G, Genchi C. Molecular evidence for a close relative of the arthropod endosymbiont Wolbachia in a filarial worm. Mol Biochem Parasitol. 1995;74:223–7. https://doi.org/10.1016/0166-6851(95)02494-8.
99. Hertig M. The Rickettsia, Wolbachia pipientis (gen. et sp.n.) and Associated Inclusions of the Mosquito, Culex pipiens. Parasitology. 1936;28:453–86. https://doi.org/10.1017/S0031182000022666.
100. Shropshire JD, Leigh B, Bordenstein SR. Symbiont-mediated cytoplasmic incompatibility: What have we learned in 50 years? eLife. 2020;9:e61989. https://doi.org/10.7554/eLife.61989.
101. Werren JH, Baldo L, Clark ME. Wolbachia: master manipulators of invertebrate biology. Nat Rev Microbiol. 2008;6:741–51. https://doi.org/10.1038/nrmicro1969.
102. Fallon AM. Growth and Maintenance of Wolbachia in Insect Cell Lines. Insects. 2021;12:706. https://doi.org/10.3390/insects12080706.
103. Turley AP, Zalucki MP, O'Neill SL, McGraw EA. Transinfected Wolbachia have minimal effects on male reproductive success in Aedes aegypti. Parasit Vectors. 2013;6:36. https://doi.org/10.1186/1756-3305-6-36.
104. Kaur R, Shropshire JD, Cross KL, Leigh B, Mansueto AJ, Stewart V, et al. Living in the endosymbiotic world of Wolbachia: A centennial review. Cell Host Microbe. 2021;29:879–93. https://doi.org/10.1016/j.chom.2021.03.006.
105. Gavotte L, Henri H, Stouthamer R, Charif D, Charlat S, Boulétreau M, et al. A Survey of the Bacteriophage WO in the Endosymbiotic Bacteria Wolbachia. Mol Biol Evol. 2007;24:427–35. https://doi.org/10.1093/molbev/msl171.
106. Bordenstein SR, Bordenstein SR. Eukaryotic association module in phage WO genomes from Wolbachia. Nat Commun. 2016;7:13155. https://doi.org/10.1038/ncomms13155.
107. Bordenstein SR, Bordenstein SR. Widespread phages of endosymbionts: Phage WO genomics and the proposed taxonomic classification of Symbioviridae. PLOS Genet. 2022;18:e1010227. https://doi.org/10.1371/journal.pgen.1010227.
108. Beckmann JF, Ronau JA, Hochstrasser M. A Wolbachia Deubiquitylating Enzyme Induces Cytoplasmic Incompatibility. Nat Microbiol. 2017;2:17007. https://doi.org/10.1038/nmicrobiol.2017.7.
109. LePage DP, Metcalf JA, Bordenstein SR, On J, Perlmutter JI, Shropshire JD, et al. Prophage WO genes recapitulate and enhance Wolbachia-induced cytoplasmic incompatibility. Nature. 2017;543:243–7. https://doi.org/10.1038/nature21391.
110. Perlmutter JI, Bordenstein SR, Unckless RL, LePage DP, Metcalf JA, Hill T, et al. The phage gene wmk is a candidate for male killing by a bacterial endosymbiont. PLOS Pathog. 2019;15:e1007936. https://doi.org/10.1371/journal.ppat.1007936.

111. Fricke LC, Lindsey ARI. Identification of Parthenogenesis-Inducing Effector Proteins in Wolbachia. Genome Biol Evol. 2024;16:evae036. https://doi.org/10.1093/gbe/evae036.
112. Grève P, Moumen B, Bouchon D. Three feminizing Wolbachia strains in a single host species: comparative genomics paves the way for identifying sex reversal factors. Front Microbiol. 2024;15. https://doi.org/10.3389/fmicb.2024.1416057.
113. Chrostek E, Teixeira L. Mutualism Breakdown by Amplification of Wolbachia Genes. PLOS Biol. 2015;13:e1002065. https://doi.org/10.1371/journal.pbio.1002065.
114. Duarte EH, Carvalho A, López-Madrigal S, Costa J, Teixeira L. Forward genetics in Wolbachia: Regulation of Wolbachia proliferation by the amplification and deletion of an addictive genomic island. PLOS Genet. 2021;17:e1009612. https://doi.org/10.1371/journal.pgen.1009612.
115. Shropshire JD, Leigh B, Bordenstein SR, Duplouy A, Riegler M, Brownlie JC, et al. Models and Nomenclature for Cytoplasmic Incompatibility: Caution over Premature Conclusions – A Response to Beckmann et al. Trends Genet. 2019;35:397–9. https://doi.org/10.1016/j.tig.2019.03.004.
116. Beckmann JF, Bonneau M, Chen H, Hochstrasser M, Poinsot D, Merçot H, et al. The Toxin–Antidote Model of Cytoplasmic Incompatibility: Genetics and Evolutionary Implications. Trends Genet. 2019;35:175–85. https://doi.org/10.1016/j.tig.2018.12.004.
117. Zhou W, Rousset F, O'Neil S. Phylogeny and PCR-based classification of Wolbachia strains using wsp gene sequences. Proc R Soc B Biol Sci. 1998;265:509–15. https://doi.org/10.1098/rspb.1998.0324.
118. Baldo L, Dunning Hotopp JC, Jolley KA, Bordenstein SR, Biber SA, Choudhury RR, et al. Multilocus Sequence Typing System for the Endosymbiont Wolbachia pipientis. Appl Environ Microbiol. 2006;72:7098–110. https://doi.org/10.1128/AEM.00731-06.
119. Schneider DI, Klasson L, Lind AE, Miller WJ. More than fishing in the dark: PCR of a dispersed sequence produces simple but ultrasensitive Wolbachia detection. BMC Microbiol. 2014;14:121. https://doi.org/10.1186/1471-2180-14-121.
120. Weinert LA, Araujo-Jnr EV, Ahmed MZ, Welch JJ. The incidence of bacterial endosymbionts in terrestrial arthropods. Proc R Soc B Biol Sci. 2015;282:20150249. https://doi.org/10.1098/rspb.2015.0249.
121. Lefoulon E, Bain O, Makepeace BL, d'Haese C, Uni S, Martin C, et al. Breakdown of coevolution between symbiotic bacteria Wolbachia and their filarial hosts. PeerJ. 2016;4:e1840. https://doi.org/10.7717/peerj.1840.
122. Balvín O, Roth S, Talbot B, Reinhardt K. Co-speciation in bedbug Wolbachia parallel the pattern in nematode hosts. Sci Rep. 2018;8:8797. https://doi.org/10.1038/s41598-018-25545-y.
123. Raychoudhury R, Baldo L, Oliveira DCSG, Werren JH. Modes of Acquisition of *Wolbachia*: Horizontal Transfer, Hybrid Introgression, and Codivergence in the *Nasonia* Species Complex. Evolution. 2008;63:165–83. https://doi.org/10.1111/j.1558-5646.2008.00533.x.
124. Gerth M, Bleidorn C. Comparative genomics provides a timeframe for Wolbachia evolution and exposes a recent biotin synthesis operon transfer. Nat Microbiol. 2016;2:16241. https://doi.org/10.1038/nmicrobiol.2016.241.
125. Bailly-Bechet M, Martins-Simões P, Szöllősi GJ, Mialdea G, Sagot M-F, Charlat S. How Long Does Wolbachia Remain on Board? Mol Biol Evol. 2017;34:1183–93. https://doi.org/10.1093/molbev/msx073.
126. Kulmuni J, Butlin RK, Lucek K, Savolainen V, Westram AM. Towards the completion of speciation: the evolution of reproductive isolation beyond the first

barriers. Philos Trans R Soc B Biol Sci. 2020;375:20190528. https://doi.org/10.1098/rstb.2019.0528.
127. Telschow A, Hammerstein P, Werren JH. The Effect of Wolbachia on Genetic Divergence between Populations: Models with Two-Way Migration. Am Nat. 2002;160:S54–66. https://doi.org/10.1086/342153.
128. Telschow A, Engelstädter J, Yamamura N, Hammerstein P, Hurst GDD. Asymmetric gene flow and constraints on adaptation caused by sex ratio distorters. J Evol Biol. 2006;19:869–78. https://doi.org/10.1111/j.1420-9101.2005.01049.x.
129. Telschow A, Hilgenboecker K, Hammerstein P, Werren JH. Dobzhansky-Muller and Wolbachia-Induced Incompatibilities in a Diploid Genetic System. PLoS ONE. 2014;9:e95488. https://doi.org/10.1371/journal.pone.0095488.
130. Gebiola M, Kelly SE, Hammerstein P, Giorgini M, Hunter MS. "Darwin's corollary" and cytoplasmic incompatibility induced by Cardinium may contribute to speciation in Encarsia wasps (Hymenoptera: Aphelinidae). Evolution. 2016;70:2447–58. https://doi.org/10.1111/evo.13037.
131. Smith DAS, Gordon IJ, Traut W, Herren J, Collins S, Martins DJ, et al. A neo-W chromosome in a tropical butterfly links colour pattern, male-killing, and speciation. Proc R Soc B Biol Sci. 2016;283:20160821. https://doi.org/10.1098/rspb.2016.0821.
132. Jaenike J, Dyer KA, Cornish C, Minhas MS. Asymmetrical Reinforcement and Wolbachia Infection in Drosophila. PLOS Biol. 2006;4:e325. https://doi.org/10.1371/journal.pbio.0040325.
133. Marshall JL. Rapid Evolution of Spermathecal Duct Length in the Allonemobius socius Complex of Crickets: Species, Population and Wolbachia Effects. PLOS ONE. 2007;2:e720. https://doi.org/10.1371/journal.pone.0000720.
134. Miller WJ, Riegler M. Evolutionary Dynamics of *w* Au-Like *Wolbachia* Variants in Neotropical *Drosophila* spp. Appl Environ Microbiol. 2006;72:826–35. https://doi.org/10.1128/AEM.72.1.826-835.2006.
135. Müller MJ, Dörr NCD, Deprá M, Schmitz HJ, Valiati VH, Valente VLDS. Reevaluating the infection status by the Wolbachia endosymbiont in Drosophila Neotropical species from the willistoni subgroup. Infect Genet Evol. 2013;19:232–9. https://doi.org/10.1016/j.meegid.2013.07.022.
136. Williamson DL, Ehrman L, Kernaghan RP. Induction of Sterility in *Drosophila paulistorum:* Effect of Cytoplasmic Factors. Proc Natl Acad Sci. 1971;68:2158–60. https://doi.org/10.1073/pnas.68.9.2158.
137. Ehrman L, Daniels S. Pole Cells of Drosophila paulistorum Embryologic Differentiation with Symbionts. Aust J Biol Sci. 1975;28:133. https://doi.org/10.1071/BI9750133.
138. Kernaghan RP, Ehrman L. Antimycoplasmal Antibiotics and Hybrid Sterility in *Drosophila paulistorum*. Science. 1970;169:63–4. https://doi.org/10.1126/science.169.3940.63.
139. Daniels S, Ehrman L. Embryonic pole cells and mycoplasmalike symbionts in *Drosophila paulistorum*. J Invertebr Pathol. 1974;24:14–9. https://doi.org/10.1016/0022-2011(74)90157-8.
140. Martinez J, Klasson L, Welch JJ, Jiggins FM. Life and Death of Selfish Genes: Comparative Genomics Reveals the Dynamic Evolution of Cytoplasmic Incompatibility. Mol Biol Evol. 2021;38:2–15. https://doi.org/10.1093/molbev/msaa209.
141. Jacobs J, Nakamoto A, Mastoras M, Loucks H, Mirchandani C, Karim L, et al. Complete de novo assembly of Wolbachia endosymbiont of Drosophila willistoni using long-read genome sequencing. Sci Rep. 2024;14:17770. https://doi.org/10.1038/s41598-024-68716-w.

142. Roger AJ, Muñoz-Gómez SA, Kamikawa R. The Origin and Diversification of Mitochondria. Curr Biol. 2017;27:R1177–92. https://doi.org/10.1016/j.cub.2017.09.015.
143. Vosseberg J, van Hooff JJE, Köstlbacher S, Panagiotou K, Tamarit D, Ettema TJG. The emerging view on the origin and early evolution of eukaryotic cells. Nature. 2024;633:295–305. https://doi.org/10.1038/s41586-024-07677-6.
144. Rath S, Sharma R, Gupta R, Ast T, Chan C, Durham TJ, et al. MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. Nucleic Acids Res. 2021;49:D1541–7. https://doi.org/10.1093/nar/gkaa1011.
145. Shtolz N, Mishmar D. The metazoan landscape of mitochondrial DNA gene order and content is shaped by selection and affects mitochondrial transcription. Commun Biol. 2023;6:93. https://doi.org/10.1038/s42003-023-04471-4.
146. Moura JP, Oliveira PJ, Urbano AM. Mitochondria: An overview of their origin, genome, architecture, and dynamics. Biochim Biophys Acta BBA - Mol Basis Dis. 2025;1871:167803. https://doi.org/10.1016/j.bbadis.2025.167803.
147. Nolfi-Donegan D, Braganza A, Shiva S. Mitochondrial electron transport chain: Oxidative phosphorylation, oxidant production, and methods of measurement. Redox Biol. 2020;37:101674. https://doi.org/10.1016/j.redox.2020.101674.
148. Rhooms S-K, Murari A, Goparaju NSV, Vilanueva M, Owusu-Ansah E. Insights from Drosophila on mitochondrial complex I. Cell Mol Life Sci CMLS. 2019;77:607–18. https://doi.org/10.1007/s00018-019-03293-0.
149. Liu Z, Lin H, Ye S, Liu Q, Meng Z, Zhang C, et al. Remarkably high activities of testicular cytochrome c in destroying reactive oxygen species and in triggering apoptosis. Proc Natl Acad Sci. 2006;103:8965–70. https://doi.org/10.1073/pnas.0603327103.
150. Lapuente-Brun E, Moreno-Loshuertos R, Acín-Pérez R, Latorre-Pellicer A, Colás C, Balsa E, et al. Supercomplex Assembly Determines Electron Flux in the Mitochondrial Electron Transport Chain. Science. 2013;340:1567–70. https://doi.org/10.1126/science.1230381.
151. Walker JE. The ATP synthase: the understood, the uncertain and the unknown. Biochem Soc Trans. 2013;41:1–16. https://doi.org/10.1042/BST20110773.
152. Kanehisa M, Furumichi M, Sato Y, Matsuura Y, Ishiguro-Watanabe M. KEGG: biological systems database as a model of the real world. Nucleic Acids Res. 2025;53:D672–7. https://doi.org/10.1093/nar/gkae909.
153. Signes A, Fernandez-Vizarra E. Assembly of mammalian oxidative phosphorylation complexes I–V and supercomplexes. Essays Biochem. 2018;62:255–70. https://doi.org/10.1042/EBC20170098.
154. Lu J, Marygold SJ, Gharib WH, Suter B. The aminoacyl-tRNA synthetases of Drosophila melanogaster. Fly (Austin). 2015;9:53–61. https://doi.org/10.1080/19336934.2015.1101196.
155. D'Souza AR, Minczuk M. Mitochondrial transcription and translation: overview. Essays Biochem. 2018;62:309–20. https://doi.org/10.1042/EBC20170102.
156. Falkenberg M, Larsson N-G, Gustafsson CM. Replication and Transcription of Human Mitochondrial DNA. Annu Rev Biochem. 2024;93:47–77. https://doi.org/10.1146/annurev-biochem-052621-092014.
157. Sloan DB, Havird JC, Sharbrough J. The on-again, off-again relationship between mitochondrial genomes and species boundaries. Mol Ecol. 2017;26:2212–36. https://doi.org/10.1111/mec.13959.
158. Lynch M, Koskella B, Schaack S. Mutation Pressure and the Evolution of Organelle Genomic Architecture. Science. 2006;311:1727–30. https://doi.org/10.1126/science.1118884.

159. Piccinini G, Iannello M, Puccio G, Plazzi F, Havird JC, Ghiselli F. Mitonuclear Coevolution, but not Nuclear Compensation, Drives Evolution of OXPHOS Complexes in Bivalves. Mol Biol Evol. 2021;38:2597–614. https://doi.org/10.1093/molbev/msab054.
160. Weaver RJ, Rabinowitz S, Thueson K, Havird JC. Genomic Signatures of Mitonuclear Coevolution in Mammals. Mol Biol Evol. 2022;39:msac233. https://doi.org/10.1093/molbev/msac233.
161. Hoekstra LA, Siddiq MA, Montooth KL. Pleiotropic Effects of a Mitochondrial–Nuclear Incompatibility Depend upon the Accelerating Effect of Temperature in Drosophila. Genetics. 2013;195:1129–39. https://doi.org/10.1534/genetics.113.154914.
162. Dobson AJ, Voigt S, Kumpitsch L, Langer L, Voigt E, Ibrahim R, et al. Mitonuclear interactions shape both direct and parental effects of diet on fitness and involve a SNP in mitoribosomal 16s rRNA. PLOS Biol. 2023;21:e3002218. https://doi.org/10.1371/journal.pbio.3002218.
163. Dobzhansky Th. A Critique of the Species Concept in Biology. Philos Sci. 1935;2:344–55.
164. Wagner WH. The Role and Taxonomic Treatment of Hybrids. BioScience. 1969;19:785–95. https://doi.org/10.2307/1294787.
165. Taylor SA, Larson EL. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. Nat Ecol Evol. 2019;3:170–7. https://doi.org/10.1038/s41559-018-0777-y.
166. Whitney KD, Randell RA, Rieseberg LH. Adaptive Introgression of Herbivore Resistance Traits in the Weedy Sunflower Helianthus annuus. 2006.
167. Martin NH, Bouck AC, Arnold ML. Detecting Adaptive Trait Introgression Between Iris fulva and I. brevicaulis in Highly Selective Field Conditions. Genetics. 2006;172:2481–9. https://doi.org/10.1534/genetics.105.053538.
168. Kim M, Cui M-L, Cubas P, Gillies A, Lee K, Chapman MA, et al. Regulatory Genes Control a Key Morphological and Ecological Trait Transferred Between Species. Science. 2008;322:1116–9. https://doi.org/10.1126/science.1164371.
169. Fu R, Zhu Y, Liu Y, Feng Y, Lu R-S, Li Y, et al. Genome-wide analyses of introgression between two sympatric Asian oak species. Nat Ecol Evol. 2022;6:924–35. https://doi.org/10.1038/s41559-022-01754-7.
170. Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, Joron M, et al. Adaptive Introgression across Species Boundaries in Heliconius Butterflies. PLOS Genet. 2012;8:e1002752. https://doi.org/10.1371/journal.pgen.1002752.
171. Herder F, Nolte AW, Pfaender J, Schwarzer J, Hadiaty RK, Schliewen UK. Adaptive radiation and hybridization in Wallace's Dreamponds: evidence from sailfin silversides in the Malili Lakes of Sulawesi. Proc R Soc B Biol Sci. 2006;273:2209–17. https://doi.org/10.1098/rspb.2006.3558.
172. Grant BR, Grant PR. Fission and fusion of Darwin's finches populations. Philos Trans R Soc B Biol Sci. 2008;363:2821–9. https://doi.org/10.1098/rstb.2008.0051.
173. Anderson TM, vonHoldt BM, Candille SI, Musiani M, Greco C, Stahler DR, et al. Molecular and Evolutionary History of Melanism in North American Gray Wolves. Science. 2009;323:1339–43. https://doi.org/10.1126/science.1165448.
174. Song Y, Endepols S, Klemann N, Richter D, Matuschka F-R, Shih C-H, et al. Adaptive Introgression of Anticoagulant Rodent Poison Resistance by Hybridization between Old World Mice. Curr Biol. 2011;21:1296–301. https://doi.org/10.1016/j.cub.2011.06.043.

175. Figueiró HV, Li G, Trindade FJ, Assis J, Pais F, Fernandes G, et al. Genome-wide signatures of complex introgression and adaptive evolution in the big cats. Sci Adv. 2017;3:e1700299. https://doi.org/10.1126/sciadv.1700299.
176. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nat Rev Genet. 2015;16:359–71. https://doi.org/10.1038/nrg3936.
177. Aguillon SM, Dodge TO, Preising GA, Schumer M. Introgression. Curr Biol. 2022;32:R865–8. https://doi.org/10.1016/j.cub.2022.07.004.
178. Brand JA, Bertram MG, Bolstad GH, Brodin T, Florencia Camus M, Hagen IJ, et al. The importance of mitochondrial DNA introgression for conservation. Trends Ecol Evol. 2026;41:775–87. https://doi.org/10.1016/j.tree.2026.04.017.
179. Lee H-Y, Chou J-Y, Cheong L, Chang N-H, Yang S-Y, Leu J-Y. Incompatibility of Nuclear and Mitochondrial Genomes Causes Hybrid Sterility between Two Yeast Species. Cell. 2008;135:1065–73. https://doi.org/10.1016/j.cell.2008.10.047.
180. Koevoets T, Niehuis O, van de Zande L, Beukeboom LW. Hybrid incompatibilities in the parasitic wasp genus Nasonia: negative effects of hemizygosity and the identification of transmission ratio distortion loci. Heredity. 2012;108:302–11. https://doi.org/10.1038/hdy.2011.75.
181. Osada N, Akashi H. Mitochondrial–Nuclear Interactions and Accelerated Compensatory Evolution: Evidence from the Primate Cytochrome c Oxidase Complex. Mol Biol Evol. 2012;29:337–46. https://doi.org/10.1093/molbev/msr211.
182. Meiklejohn CD, Holmbeck MA, Siddiq MA, Abt DN, Rand DM, Montooth KL. An Incompatibility between a Mitochondrial tRNA and Its Nuclear-Encoded tRNA Synthetase Compromises Development and Fitness in Drosophila. PLOS Genet. 2013;9:e1003238. https://doi.org/10.1371/journal.pgen.1003238.
183. Chang C-C, Rodriguez J, Ross J. Mitochondrial–Nuclear Epistasis Impacts Fitness and Mitochondrial Physiology of Interpopulation Caenorhabditis briggsae Hybrids. G3 GenesGenomesGenetics. 2016;6:209–19. https://doi.org/10.1534/g3.115.022970.
184. Morales HE, Pavlova A, Amos N, Major R, Kilian A, Greening C, et al. Concordant divergence of mitogenomes and a mitonuclear gene cluster in bird lineages inhabiting different climates. Nat Ecol Evol. 2018;2:1258–67. https://doi.org/10.1038/s41559-018-0606-3.
185. Lima TG, Burton RS, Willett CS. Genomic scans reveal multiple mito-nuclear incompatibilities in population crosses of the copepod Tigriopus californicus. Evolution. 2019;73:609–20. https://doi.org/10.1111/evo.13690.
186. Barth JMI, Gubili C, Matschiner M, Tørresen OK, Watanabe S, Egger B, et al. Stable species boundaries despite ten million years of hybridization in tropical eels. Nat Commun. 2020;11:1433. https://doi.org/10.1038/s41467-020-15099-x.
187. McKenzie JL, Araújo HA, Smith JL, Schluter D, Devlin RH. Incomplete reproductive isolation and strong transcriptomic response to hybridization between sympatric sister species of salmon. Proc R Soc B Biol Sci. 2021;288:20203020. https://doi.org/10.1098/rspb.2020.3020.
188. Sato Y, Fujiwara S, Egas M, Matsuda T, Gotoh T. Patterns of reproductive isolation in a haplodiploid mite, Amphitetranychus viennensis: prezygotic isolation, hybrid inviability and hybrid sterility. BMC Ecol Evol. 2021;21:177. https://doi.org/10.1186/s12862-021-01896-5.
189. Healy TM, Burton RS. Differential gene expression and mitonuclear incompatibilities in fast- and slow-developing interpopulation Tigriopus californicus hybrids. Mol Ecol. 2023;32:3102–17. https://doi.org/10.1111/mec.16917.

190. Moran BM, Payne CY, Powell DL, Iverson ENK, Donny AE, Banerjee SM, et al. A lethal mitonuclear incompatibility in complex I of natural hybrids. Nature. 2024;626:119–27. https://doi.org/10.1038/s41586-023-06895-8.
191. Beck EA, Thompson AC, Sharbrough J, Brud E, Llopart A. Gene flow between Drosophila yakuba and Drosophila santomea in subunit V of cytochrome c oxidase: A potential case of cytonuclear cointrogression. Evolution. 2015;69:1973–86. https://doi.org/10.1111/evo.12718.
192. Boratyński Z, Ketola T, Koskela E, Mappes T. The Sex Specific Genetic Variation of Energetics in Bank Voles, Consequences of Introgression? Evol Biol. 2016;43:37–47. https://doi.org/10.1007/s11692-015-9347-2.
193. Seixas FA, Boursot P, Melo-Ferreira J. The genomic impact of historical hybridization with massive mitochondrial DNA introgression. Genome Biol. 2018;19:91. https://doi.org/10.1186/s13059-018-1471-8.
194. Wang S, Ore MJ, Mikkelsen EK, Lee-Yaw J, Toews DPL, Rohwer S, et al. Signatures of mitonuclear coevolution in a warbler species complex. Nat Commun. 2021;12:4279. https://doi.org/10.1038/s41467-021-24586-8.
195. Nikelski E, Rubtsov AS, Irwin D. High heterogeneity in genomic differentiation between phenotypically divergent songbirds: a test of mitonuclear co-introgression. Heredity. 2023;130:1–13. https://doi.org/10.1038/s41437-022-00580-8.
196. Pettersson ME, Fuentes-Pardo AP, Rochus CM, Enbody ED, Bi H, Väinölä R, et al. A Long-Standing Hybrid Population Between Pacific and Atlantic Herring in a Subarctic Fjord of Norway. Genome Biol Evol. 2023;15:evad069. https://doi.org/10.1093/gbe/evad069.
197. Bailey NP, Stevison LS. Mitonuclear conflict in a macaque species exhibiting phylogenomic discordance. J Evol Biol. 2021;34:1568–79. https://doi.org/10.1111/jeb.13914.
198. Evans BJ, Peter BM, Melnick DJ, Andayani N, Supriatna J, Zhu J, et al. Mitonuclear interactions and introgression genomics of macaque monkeys (Macaca) highlight the influence of behaviour on genome evolution. Proc R Soc B Biol Sci. 2021;288:20211756. https://doi.org/10.1098/rspb.2021.1756.
199. Biot-Pelletier D, Bettinazzi S, Gagnon-Arsenault I, Dubé AK, Bédard C, Nguyen THM, et al. Evolutionary Trajectories are Contingent on Mitonuclear Interactions. Mol Biol Evol. 2023;40:msad061. https://doi.org/10.1093/molbev/msad061.
200. Doiron S, Bernatchez L, Blier PU. A Comparative Mitogenomic Analysis of the Potential Adaptive Value of Arctic Charr mtDNA Introgression in Brook Charr Populations (Salvelinus fontinalis Mitchill). Mol Biol Evol. 2002;19:1902–9. https://doi.org/10.1093/oxfordjournals.molbev.a004014.
201. Boratyński Z, Melo-Ferreira J, Alves PC, Berto S, Koskela E, Pentikäinen OT, et al. Molecular and ecological signs of mitochondrial adaptation: consequences for introgression? Heredity. 2014;113:277–86. https://doi.org/10.1038/hdy.2014.28.
202. Melo-Ferreira J, Vilela J, Fonseca MM, da Fonseca RR, Boursot P, Alves PC. The Elusive Nature of Adaptive Mitochondrial DNA Evolution of an Arctic Lineage Prone to Frequent Introgression. Genome Biol Evol. 2014;6:886–96. https://doi.org/10.1093/gbe/evu059.
203. Llopart A, Herrig D, Brud E, Stecklein Z. Sequential adaptive introgression of the mitochondrial genome in Drosophila yakuba and Drosophila santomea. Mol Ecol. 2014;23:1124–36. https://doi.org/10.1111/mec.12678.
204. Currat M, Ruedi M, Petit RJ, Excoffier L. The Hidden Side of Invasions: Massive Introgression by Local Genes. Evolution. 2008;62:1908–20. https://doi.org/10.1111/j.1558-5646.2008.00413.x.

205. Mastrantonio V, Porretta D, Urbanelli S, Crasta G, Nascetti G. Dynamics of mtDNA introgression during species range expansion: insights from an experimental longitudinal study. Sci Rep. 2016;6:30355. https://doi.org/10.1038/srep30355.
206. Turelli M, Hoffmann AA, McKechnie SW. Dynamics of Cytoplasmic Incompatibility and Mtdna Variation in Natural Drosophila Simulans Populations. Genetics. 1992;132:713–23. https://doi.org/10.1093/genetics/132.3.713.
207. Jiggins FM. Male-Killing Wolbachia and Mitochondrial DNA: Selective Sweeps, Hybrid Introgression and Parasite Population Dynamics. Genetics. 2003;164:5–12. https://doi.org/10.1093/genetics/164.1.5.
208. Hurst GDD, Jiggins FM. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proc R Soc B Biol Sci. 2005;272:1525–34. https://doi.org/10.1098/rspb.2005.3056.
209. Willis SC, Farias IP, Ortí G. TESTING MITOCHONDRIAL CAPTURE AND DEEP COALESCENCE IN AMAZONIAN CICHLID FISHES (CICHLIDAE: CICHLA). Evolution. 2014;68:256–68. https://doi.org/10.1111/evo.12230.
210. Mikkelsen EK, Weir JT. Phylogenomics Reveals that Mitochondrial Capture and Nuclear Introgression Characterize Skua Species Proposed to be of Hybrid Origin. Syst Biol. 2023;72:78–91. https://doi.org/10.1093/sysbio/syac078.
211. Hibbins MS, Hahn MW. Phylogenomic approaches to detecting and characterizing introgression. Genetics. 2022;220:iyab173. https://doi.org/10.1093/genetics/iyab173.
212. Maddison WP. Gene Trees in Species Trees. Syst Biol. 1997;46:523–36. https://doi.org/10.1093/sysbio/46.3.523.
213. Degnan JH, Rosenberg NA. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol. 2009;24:332–40. https://doi.org/10.1016/j.tree.2009.01.009.
214. Edwards ScottV, Beerli P. PERSPECTIVE: GENE DIVERGENCE, POPULATION DIVERGENCE, AND THE VARIANCE IN COALESCENCE TIME IN PHYLOGEOGRAPHIC STUDIES. Evolution. 2000;54:1839–54. https://doi.org/10.1111/j.0014-3820.2000.tb01231.x.
215. Montoliu-Nerin M, Strunov A, Heyworth E, Schneider DI, Thoma J, Hua-Van A, et al. Evolutionary persistence of a highly prevalent multicopy mitochondrial-derived nuclear insertion (Mega-NUMT) in Neotropical Drosophila flies. Mol Biol Evol. 2026;:msag227. https://doi.org/10.1093/molbev/msag227.
216. Richly E, Leister D. NUMTs in Sequenced Eukaryotic Genomes. Mol Biol Evol. 2004;21:1081–4. https://doi.org/10.1093/molbev/msh110.
217. Hazkani-Covo E, Zeller RM, Martin W. Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. PLOS Genet. 2010;6:e1000834. https://doi.org/10.1371/journal.pgen.1000834.
218. Zhou W, Karan KR, Gu W, Klein H-U, Sturm G, Jager PLD, et al. Somatic nuclear mitochondrial DNA insertions are prevalent in the human brain and accumulate over time in fibroblasts. PLOS Biol. 2024;22:e3002723. https://doi.org/10.1371/journal.pbio.3002723.
219. Srinivasainagendra V, Sandel MW, Singh B, Sundaresan A, Mooga VP, Bajpai P, et al. Migration of mitochondrial DNA in the nuclear genome of colorectal adenocarcinoma. Genome Med. 2017;9:31. https://doi.org/10.1186/s13073-017-0420-6.
220. Thorsness PE, Fox TD. Escape of DNA from mitochondria to the nucleus in Saccharomyces cerevisiae. Nature. 1990;346:376–9. https://doi.org/10.1038/346376a0.

221. Dayama G, Zhou W, Prado-Martinez J, Marques-Bonet T, Mills RE. Characterization of nuclear mitochondrial insertions in the whole genomes of primates. NAR Genomics Bioinforma. 2020;2:lqaa089. https://doi.org/10.1093/nargab/lqaa089.
222. Wei W, Schon KR, Elgar G, Orioli A, Tanguy M, Giess A, et al. Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. Nature. 2022;611:105–14. https://doi.org/10.1038/s41586-022-05288-7.
223. Zhu Q, Zhang J, Zhou W, Liang S-A, Wang S, Cai X, et al. Introgressed mitochondrial fragments from archaic hominins alter nuclear genome function in modern humans. Sci Adv. 2026;12:eaea0706. https://doi.org/10.1126/sciadv.aea0706.
224. Wells JN, Feschotte C. A Field Guide to Eukaryotic Transposable Elements. Annu Rev Genet. 2020;54 Volume 54, 2020:539–61. https://doi.org/10.1146/annurev-genet-040620-022145.
225. Nugent JM, Palmer JD. RNA-mediated transfer of the gene *coxII* from the mitochondrion to the nucleus during flowering plant evolution. Cell. 1991;66:473–81. https://doi.org/10.1016/0092-8674(81)90011-8.
226. Adams KL, Song K, Roessler PG, Nugent JM, Doyle JL, Doyle JJ, et al. Intracellular gene transfer in action: Dual transcription and multiple silencings of nuclear and mitochondrial cox2 genes in legumes. Proc Natl Acad Sci. 1999;96:13863–8. https://doi.org/10.1073/pnas.96.24.13863.
227. Daley DO, Adams KL, Clifton R, Qualmann S, Millar AH, Palmer JD, et al. Gene transfer from mitochondrion to nucleus: novel mechanisms for gene activation from Cox2. Plant J. 2002;30:11–21. https://doi.org/10.1046/j.1365-313X.2002.01263.x.
228. Alverson AJ, Zhuo S, Rice DW, Sloan DB, Palmer JD. The Mitochondrial Genome of the Legume Vigna radiata and the Analysis of Recombination across Short Mitochondrial Repeats. PLoS ONE. 2011;6:e16404. https://doi.org/10.1371/journal.pone.0016404.
229. Bi C, Lu N, Xu Y, He C, Lu Z. Characterization and Analysis of the Mitochondrial Genome of Common Bean (Phaseolus vulgaris) by Comparative Genomic Approaches. Int J Mol Sci. 2020;21:3778. https://doi.org/10.3390/ijms21113778.
230. Covello PS, Gray MW. Silent mitochondrial and active nuclear genes for subunit 2 of cytochrome c oxidase (cox2) in soybean: evidence for RNA-mediated gene transfer. EMBO J. 1992;11:3815–20. https://doi.org/10.1002/j.1460-2075.1992.tb05473.x.
231. Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, et al. Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: Implication of potential sequencing errors caused by large-unit repeats. Proc Natl Acad Sci. 2001;98:5099–103. https://doi.org/10.1073/pnas.091110398.
232. Shafer KS, Hanekamp T, White KH, Thorsness PE. Mechanisms of mitochondrial DNA escape to the nucleus in the yeast Saccharomyces cerevisiae. Curr Genet. 1999;36:183–94. https://doi.org/10.1007/s002940050489.
233. Yu Y, Wang X, Fox J, Yu R, Thakre P, McCauley B, et al. Yeast EndoG prevents genome instability by degrading extranuclear DNA species. Nat Commun. 2024;15:7653. https://doi.org/10.1038/s41467-024-52147-2.
234. Ricchetti M, Fairhead C, Dujon B. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. Nature. 1999;402:96–100. https://doi.org/10.1038/47076.
235. Singh KK, Choudhury AR, Tiwari HK. Numtogenesis as a mechanism for development of cancer. Semin Cancer Biol. 2017;47:101–9. https://doi.org/10.1016/j.semcancer.2017.05.003.

236. Thorsness PE, Fox TD. Nuclear Mutations in Saccharomyces Cerevisiae That Affect the Escape of DNA from Mitochondria to the Nucleus. Genetics. 1993;134:21–8. https://doi.org/10.1093/genetics/134.1.21.
237. Kan KT, Wilcock J, Lu H. Role of Yme1 in mitochondrial protein homeostasis: from regulation of protein import, OXPHOS function to lipid synthesis and mitochondrial dynamics. Biochem Soc Trans. 2024;52:1539–48. https://doi.org/10.1042/BST20240450.
238. Campbell CL, Thorsness PE. Escape of mitochondrial DNA to the nucleus in yme1 yeast is mediated by vacuolar-dependent turnover of abnormal mitochondrial compartments. J Cell Sci. 1998;111 (Pt 16):2455–64. https://doi.org/10.1242/jcs.111.16.2455.
239. Cupo RR, Domínguez-Martín E, Youle R. The mitochondrial intermembrane space nuclease Nuc1 (endonuclease G) prevents mitophagy-mediated mtDNA escape in yeast. Curr Biol. 2026;36:912-927.e4. https://doi.org/10.1016/j.cub.2026.01.006.
240. Sfeir A, Symington LS. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? Trends Biochem Sci. 2015;40:701–14. https://doi.org/10.1016/j.tibs.2015.08.006.
241. Ricchetti M, Tekaia F, Dujon B. Continued Colonization of the Human Genome by Mitochondrial DNA. PLOS Biol. 2004;2:e273. https://doi.org/10.1371/journal.pbio.0020273.
242. Fu L, Chen J, Lian D, Du S, Wu D, Yang C, et al. A long-read human pangenome initiative for comprehensive interpretation of nuclear-embedded mitochondrial DNA. Nat Commun. 2026;17:4371. https://doi.org/10.1038/s41467-026-71348-5.
243. Tsuji J, Frith MC, Tomii K, Horton P. Mammalian NUMT insertion is non-random. Nucleic Acids Res. 2012;40:9073–88. https://doi.org/10.1093/nar/gks424.
244. Behura SK. Analysis of Nuclear Copies of Mitochondrial Sequences in Honeybee (Apis mellifera) Genome. Mol Biol Evol. 2007;24:1492–505. https://doi.org/10.1093/molbev/msm068.
245. Mourier T, Hansen AJ, Willerslev E, Arctander P. The Human Genome Project Reveals a Continuous Transfer of Large Mitochondrial Fragments to the Nucleus. Mol Biol Evol. 2001;18:1833–7. https://doi.org/10.1093/oxfordjournals.molbev.a003971.
246. Hallberg BM, Larsson N-G. TFAM forces mtDNA to make a U-turn. Nat Struct Mol Biol. 2011;18:1179–81. https://doi.org/10.1038/nsmb.2167.
247. Antil S, Abraham JS, Sripoorna S, Maurya S, Dagar J, Makhija S, et al. DNA barcoding, an effective tool for species identification: a review. Mol Biol Rep. 2023;50:761–75. https://doi.org/10.1007/s11033-022-08015-7.
248. Colombo WD, de Freitas Justino J, Barcelos AC, Vilaça ST, Pavanelli L, Vargas SM. Reassessing leatherback turtle lineages and unveiling the first evidence of nuclear mitochondrial DNA in sea turtles. Sci Rep. 2024;14:31313. https://doi.org/10.1038/s41598-024-82754-4.
249. Balciuniene J, Balciunas D. A Nuclear mtDNA Concatemer (Mega-NUMT) Could Mimic Paternal Inheritance of Mitochondrial Genome. Front Genet. 2019;10. https://doi.org/10.3389/fgene.2019.00518.
250. Wei W, Pagnamenta AT, Gleadall N, Sanchis-Juan A, Stephens J, Broxholme J, et al. Nuclear-mitochondrial DNA segments resemble paternally inherited mitochondrial DNA in humans. Nat Commun. 2020;11:1740. https://doi.org/10.1038/s41467-020-15336-3.
251. Ghezzi D, Zeviani M. Human diseases associated with defects in assembly of OXPHOS complexes. Essays Biochem. 2018;62:271–86. https://doi.org/10.1042/EBC20170099.

252. Frazier AE, Thorburn DR, Compton AG. Mitochondrial energy generation disorders: genes, mechanisms, and clues to pathology. J Biol Chem. 2019;294:5386–95. https://doi.org/10.1074/jbc.R117.809194.
253. Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J Mol Evol. 1994;39:174–90. https://doi.org/10.1007/BF00163806.
254. Lutz-Bonengel S, Niederstätter H, Naue J, Koziel R, Yang F, Sänger T, et al. Evidence for multi-copy Mega-NUMTs in the human genome. Nucleic Acids Res. 2021;49:1517–31. https://doi.org/10.1093/nar/gkaa1271.
255. Bzymek M, Lovett ST. Instability of repetitive DNA sequences: The role of replication in multiple mechanisms. Proc Natl Acad Sci. 2001;98:8319–25. https://doi.org/10.1073/pnas.111008398.
256. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev Genet. 2009;10:551–64. https://doi.org/10.1038/nrg2593.
257. Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res. 2008;36:6688–719. https://doi.org/10.1093/nar/gkn668.
258. Kuzminov A. The Precarious Prokaryotic Chromosome. J Bacteriol. 2014;196:1793–806. https://doi.org/10.1128/JB.00022-14.
259. Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philos Trans R Soc B Biol Sci. 2015;370:20140331. https://doi.org/10.1098/rstb.2014.0331.
260. Jumas-Bilak E, Michaux-Charachon S, Bourg G, O'Callaghan D, Ramuz M. Differences in chromosome number and genome rearrangements in the genus Brucella. Mol Microbiol. 1998;27:99–106. https://doi.org/10.1046/j.1365-2958.1998.00661.x.
261. Yamaichi Y, Iida T, Park K-S, Yamamoto K, Honda T. Physical and genetic map of the genome of Vibrio parahaemolyticus: presence of two chromosomes in Vibrio species. Mol Microbiol. 1999;31:1513–21. https://doi.org/10.1046/j.1365-2958.1999.01296.x.
262. Baliga NS, Bonneau R, Facciotti MT, Pan M, Glusman G, Deutsch EW, et al. Genome sequence of Haloarcula marismortui: A halophilic archaeon from the Dead Sea. Genome Res. 2004;14:2221–34. https://doi.org/10.1101/gr.2700304.
263. Volff J-N, Altenbuchner J. A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. FEMS Microbiol Lett. 2000;186:143–50. https://doi.org/10.1111/j.1574-6968.2000.tb09095.x.
264. Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, et al. Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature. 2001;414:450–3. https://doi.org/10.1038/35106579.
265. Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, et al. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucleic Acids Res. 2002;30:4264–71. https://doi.org/10.1093/nar/gkf549.
266. Lee Y, Rio DC. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annu Rev Biochem. 2015;84:291–323. https://doi.org/10.1146/annurev-biochem-060614-034316.
267. Koonin EV. Evolution of genome architecture. Int J Biochem Cell Biol. 2009;41:298–306. https://doi.org/10.1016/j.biocel.2008.09.015.
268. Touchon M, Rocha EPC. Causes of Insertion Sequences Abundance in Prokaryotic Genomes. Mol Biol Evol. 2007;24:969–81. https://doi.org/10.1093/molbev/msm014.

269. Mrázek J, Guo X, Shah A. Simple sequence repeats in prokaryotic genomes. Proc Natl Acad Sci. 2007;104:8472–7. https://doi.org/10.1073/pnas.0702412104.
270. Garrido-Ramos MA, Plohl M, Šatović-Vukšić E. Satellite DNA Genomics: The Ongoing Story. Int J Mol Sci. 2025;26:11291. https://doi.org/10.3390/ijms262311291.
271. Fuentes-Pardo AP, Ruzzante DE. Whole-genome sequencing approaches for conservation biology: Advantages, limitations and practical recommendations. Mol Ecol. 2017;26:5369–406. https://doi.org/10.1111/mec.14264.
272. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 2019;20:92. https://doi.org/10.1186/s13059-019-1715-2.
273. O'Leary NA, Cox E, Holmes JB, Anderson WR, Falk R, Hem V, et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. Sci Data. 2024;11:732. https://doi.org/10.1038/s41597-024-03571-y.
274. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977;74:5463–7. https://doi.org/10.1073/pnas.74.12.5463.
275. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921. https://doi.org/10.1038/35057062.
276. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26:1135–45. https://doi.org/10.1038/nbt1486.
277. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51. https://doi.org/10.1038/nrg.2016.49.
278. Hess JF, Kohl TA, Kotrová M, Rönsch K, Paprotka T, Mohr V, et al. Library preparation for next generation sequencing: A review of automation strategies. Biotechnol Adv. 2020;41:107537. https://doi.org/10.1016/j.biotechadv.2020.107537.
279. Risca VI, Greenleaf illiam J. Beyond the Linear Genome: Paired-End Sequencing as a Biophysical Tool. Trends Cell Biol. 2015;25:716–9. https://doi.org/10.1016/j.tcb.2015.08.004.
280. Liu T, Chen C-Y, Chen-Deng A, Chen Y-L, Wang J-Y, Hou Y-I, et al. Joining Illumina paired-end reads for classifying phylogenetic marker sequences. BMC Bioinformatics. 2020;21:105. https://doi.org/10.1186/s12859-020-3445-6.
281. Segerman B. The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. Front Cell Infect Microbiol. 2020;10. https://doi.org/10.3389/fcimb.2020.527102.
282. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet. 2018;34:666–81. https://doi.org/10.1016/j.tig.2018.05.008.
283. Espinosa E, Bautista R, Larrosa R, Plata O. Advancements in long-read genome sequencing technologies and algorithms. Genomics. 2024;116:110842. https://doi.org/10.1016/j.ygeno.2024.110842.
284. Sohn J, Nam J-W. The present and future of de novo whole-genome assembly. Brief Bioinform. 2018;19:23–40. https://doi.org/10.1093/bib/bbw096.
285. Ferrarini M, Moretto M, Ward JA, Šurbanovski N, Stevanović V, Giongo L, et al. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. BMC Genomics. 2013;14:670. https://doi.org/10.1186/1471-2164-14-670.

286. Schwartz S, Oren R, Ast G. Detection and Removal of Biases in the Analysis of Next-Generation Sequencing Reads. PLOS ONE. 2011;6:e16685. https://doi.org/10.1371/journal.pone.0016685.
287. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40:e72. https://doi.org/10.1093/nar/gks001.
288. Delahaye C, Nicolas J. Sequencing DNA with nanopores: Troubles and biases. PLOS ONE. 2021;16:e0257521. https://doi.org/10.1371/journal.pone.0257521.
289. Carvalho AB, Kim BY, Uno F. Strong bias in long-read sequencing prevents assembly of Drosophila melanogaster Y-linked genes. Genome Res. 2026;36:71–82. https://doi.org/10.1101/gr.280604.125.
290. Chen Z, Ong CT, Nguyen LT, Lamb HJ, González-Recio O, Gutiérrez-Rivas M, et al. Biases from Oxford Nanopore library preparation kits and their effects on microbiome and genome analysis. BMC Genomics. 2025;26:504. https://doi.org/10.1186/s12864-025-11649-z.
291. Winden E, Vasquez-Echeverri A, Calle-Castañeda S, Lian Y, Ortiz JPH, Schwartz DC, et al. A database of restriction maps to expand the utility of bacterial artificial chromosomes. Gigabyte. 2023;2023:1–5. https://doi.org/10.46471/gigabyte.93.
292. Taylor DJ, Eizenga JM, Li Q, Das A, Jenike KM, Kenny EE, et al. Beyond the Human Genome Project: The Age of Complete Human Genome Sequences and Pangenome References. Annu Rev Genomics Hum Genet. 2024;25 Volume 25, 2024:77–104. https://doi.org/10.1146/annurev-genom-021623-081639.
293. Li H, Durbin R. Genome assembly in the telomere-to-telomere era. Nat Rev Genet. 2024;25:658–70. https://doi.org/10.1038/s41576-024-00718-w.
294. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol J Comput Mol Cell Biol. 2012;19:455–77. https://doi.org/10.1089/cmb.2012.0021.
295. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36. https://doi.org/10.1101/gr.215087.116.
296. Ekblom R, Wolf JBW. A field guide to whole-genome sequencing, assembly and annotation. Evol Appl. 2014;7:1026–42. https://doi.org/10.1111/eva.12178.
297. Luo J, Wei Y, Lyu M, Wu Z, Liu X, Luo H, et al. A comprehensive review of scaffolding methods in genome assembly. Brief Bioinform. 2021;22:bbab033. https://doi.org/10.1093/bib/bbab033.
298. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011;8:61–5. https://doi.org/10.1038/nmeth.1527.
299. Trigodet F, Sachdeva R, Banfield JF, Eren AM. Troubleshooting common errors in assemblies of long-read metagenomes. Nat Biotechnol. 2026;:1–10. https://doi.org/10.1038/s41587-025-02971-8.
300. Pryszcz LP, Gabaldón T. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 2016;44:e113. https://doi.org/10.1093/nar/gkw294.
301. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19:460. https://doi.org/10.1186/s12859-018-2485-7.
302. Sun Q, Li Y. Advances in haplotype phasing and genotype imputation. Nat Rev Genet. 2026;27:155–69. https://doi.org/10.1038/s41576-025-00895-2.
303. Goldberg JK, Allan CW, Copetti D, Matzkin LM, Bronstein J. A pooled-sample draft genome assembly provides insights into host plant-specific transcriptional responses of a Solanaceae-specializing pest, Tupiocoris notatus (Hemiptera: Miridae). Ecol Evol. 2024;14:e10979. https://doi.org/10.1002/ece3.10979.

304. Powell JR, Evans BR. How Much Does Inbreeding Reduce Heterozygosity? Empirical Results from Aedes aegypti. Am J Trop Med Hyg. 2017;96:157–8. https://doi.org/10.4269/ajtmh.16-0693.
305. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13:329–42. https://doi.org/10.1038/nrg3174.
306. Richardson EJ, Watson M. The automatic annotation of bacterial genomes. Brief Bioinform. 2013;14:1–12. https://doi.org/10.1093/bib/bbs007.
307. Öztürk-Çolak A, Marygold SJ, Antonazzo G, Attrill H, Goutte-Gattat D, Jenkins VK, et al. FlyBase: updates to the Drosophila genes and genomes database. Genetics. 2024;227:iyad211. https://doi.org/10.1093/genetics/iyad211.
308. Dyer SC, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Barrera-Enriquez VP, et al. Ensembl 2025. Nucleic Acids Res. 2025;53:D948–57. https://doi.org/10.1093/nar/gkae1071.
309. Karp PD, Paley S, Caspi R, Kothari A, Krummenacker M, Midford PE, et al. The EcoCyc database (2025). EcoSal Plus. 2025;13:eesp-0019-2024. https://doi.org/10.1128/ecosalplus.esp-0019-2024.
310. Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 2001;29:2607–18. https://doi.org/10.1093/nar/29.12.2607.
311. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34 suppl_2:W435–9. https://doi.org/10.1093/nar/gkl200.
312. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119. https://doi.org/10.1186/1471-2105-11-119.
313. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44. https://doi.org/10.1093/bioinformatics/btn013.
314. Brůna T, Lomsadze A, Borodovsky M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. Genome Res. 2024;34:757–68. https://doi.org/10.1101/gr.278373.123.
315. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9. https://doi.org/10.1093/bioinformatics/btu153.
316. Poelzer J, Han S, Saha S, Oler E, Kruger R, Berjanskii M, et al. BASys2: a next-generation bacterial genome annotation system. Nucleic Acids Res. 2025;53:W57–67. https://doi.org/10.1093/nar/gkaf360.
317. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44:6614–24. https://doi.org/10.1093/nar/gkw569.
318. Campbell MS, Holt C, Moore B, Yandell M. Genome Annotation and Curation Using MAKER and MAKER-P. Curr Protoc Bioinforma. 2014;48:4.11.1-4.11.39. https://doi.org/10.1002/0471250953.bi0411s48.
319. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res. 2024;34:769–77. https://doi.org/10.1101/gr.278090.123.
320. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9:R7. https://doi.org/10.1186/gb-2008-9-1-r7.

321. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014;42:e119. https://doi.org/10.1093/nar/gku557.
322. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinforma. 2021;3:lqaa108. https://doi.org/10.1093/nargab/lqaa108.
323. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32:767–9. https://doi.org/10.1093/bioinformatics/btv661.
324. de Crécy-lagard V, Amorin de Hegedus R, Arighi C, Babor J, Bateman A, Blaby I, et al. A roadmap for the functional annotation of protein families: a community perspective. Database. 2022;2022:baac062. https://doi.org/10.1093/database/baac062.
325. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309–38. https://doi.org/10.1146/annurev.genet.39.073003.114725.
326. Ponting CP. Issues in predicting protein function from sequence. Brief Bioinform. 2001;2:19–29. https://doi.org/10.1093/bib/2.1.19.
327. Tekaia F. Inferring Orthologs: Open Questions and Perspectives. Genomics Insights. 2016;9:17–28. https://doi.org/10.4137/GEI.S37925.
328. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. PLOS Comput Biol. 2009;5:e1000605. https://doi.org/10.1371/journal.pcbi.1000605.
329. de Crécy-Lagard V. Variations in metabolic pathways create challenges for automated metabolic reconstructions: Examples from the tetrahydrofolate synthesis pathway. Comput Struct Biotechnol J. 2014;10:41–50. https://doi.org/10.1016/j.csbj.2014.05.008.
330. Deutsch EW, Lindskog C, Budnik B, Ctortecka C, Guo T, Pineau C, et al. The 2025 Report on the Human Proteome from the HUPO Human Proteome Project. J Proteome Res. 2026;25:539–55. https://doi.org/10.1021/acs.jproteome.5c00759.
331. Mohr SE, Kim A-R, Hu Y, Perrimon N. Finding information about uncharacterized Drosophila melanogaster genes. Genetics. 2023;225:iyad187. https://doi.org/10.1093/genetics/iyad187.
332. Reiser L, Bakker E, Subramaniam S, Chen X, Sawant S, Khosa K, et al. The Arabidopsis Information Resource in 2024. Genetics. 2024;227:iyae027. https://doi.org/10.1093/genetics/iyae027.
333. Chakraborty S, Ardern Z, Aliyu H, Kaster A-K. Deciphering the proteome of Escherichia coli K-12: Integrating transcriptomics and machine learning to annotate hypothetical proteins. Comput Struct Biotechnol J. 2025;27:3565–78. https://doi.org/10.1016/j.csbj.2025.07.036.
334. Lobb B, Tremblay BJ-M, Moreno-Hagelsieb G, Doxey AC. An assessment of genome annotation coverage across the bacterial tree of life. Microb Genomics. 2020;6:e000341. https://doi.org/10.1099/mgen.0.000341.
335. Meadows JRS, Lindblad-Toh K. Dissecting evolution and disease using comparative vertebrate genomics. Nat Rev Genet. 2017;18:624–36. https://doi.org/10.1038/nrg.2017.51.
336. Darwin C. On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life. John Murray; 1859.

337. Theobald DL. A formal test of the theory of universal common ancestry. Nature. 2010;465:219–22. https://doi.org/10.1038/nature09014.
338. Weiss MC, Sousa FL, Mrnjavac N, Neukirchen S, Roettger M, Nelson-Sathi S, et al. The physiology and habitat of the last universal common ancestor. Nat Microbiol. 2016;1:16116. https://doi.org/10.1038/nmicrobiol.2016.116.
339. Slatkin M. Gene Flow and the Geographic Structure of Natural Populations. Science. 1987;236:787–92. https://doi.org/10.1126/science.3576198.
340. Tigano A, Friesen VL. Genomics of local adaptation with gene flow. Mol Ecol. 2016;25:2144–64. https://doi.org/10.1111/mec.13606.
341. Kosiol C, Vinař T, Fonseca RR da, Hubisz MJ, Bustamante CD, Nielsen R, et al. Patterns of Positive Selection in Six Mammalian Genomes. PLOS Genet. 2008;4:e1000144. https://doi.org/10.1371/journal.pgen.1000144.
342. Bergeron LA, Besenbacher S, Zheng J, Li P, Bertelsen MF, Quintard B, et al. Evolution of the germline mutation rate across vertebrates. Nature. 2023;615:285–91. https://doi.org/10.1038/s41586-023-05752-y.
343. Dasmahapatra KK, Walters JR, Briscoe AD, Davey JW, Whibley A, Nadeau NJ, et al. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature. 2012;487:94–8. https://doi.org/10.1038/nature11041.
344. Hardison RC. Comparative Genomics. PLOS Biol. 2003;1:e58. https://doi.org/10.1371/journal.pbio.0000058.
345. Beckmann JF, Fallon AM. Detection of the *Wolbachia* protein WPIP0282 in mosquito spermathecae: Implications for cytoplasmic incompatibility. Insect Biochem Mol Biol. 2013;43:867–78. https://doi.org/10.1016/j.ibmb.2013.07.002.
346. Lamichhaney S, Berglund J, Almén MS, Maqbool K, Grabherr M, Martinez-Barrio A, et al. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature. 2015;518:371–5. https://doi.org/10.1038/nature14181.
347. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005;438:803–19. https://doi.org/10.1038/nature04338.
348. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 2003;423:241–54. https://doi.org/10.1038/nature01644.
349. Kriventseva EV, Rahman N, Espinosa O, Zdobnov EM. OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Res. 2008;36 Database issue:D271–5. https://doi.org/10.1093/nar/gkm845.
350. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Res. 2003;13:2178–89. https://doi.org/10.1101/gr.1224503.
351. Zheng XH, Lu F, Wang Z-Y, Zhong F, Hoover J, Mural R. Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. Bioinformatics. 2005;21:703–10. https://doi.org/10.1093/bioinformatics/bti045.
352. Coombe L, Kazemi P, Wong J, Birol I, Warren RL. ntSynt: multi-genome synteny detection using minimizer graph mappings. BMC Biol. 2025;23:367. https://doi.org/10.1186/s12915-025-02455-w.
353. Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A. Evolutionary constraints on structural similarity in orthologs and paralogs. Protein Sci Publ Protein Soc. 2009;18:1306–15. https://doi.org/10.1002/pro.143.
354. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma Oxf Engl. 2015;31:3210–2. https://doi.org/10.1093/bioinformatics/btv351.

355. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:245. https://doi.org/10.1186/s13059-020-02134-9.
356. Hoffmann AA, Clancy D, Duncan J. Naturally-occurring Wolbachia infection in Drosophila simulans that does not cause cytoplasmic incompatibility. Heredity. 1996;76:1–8. https://doi.org/10.1038/hdy.1996.1.
357. Zanini R, Müller MJ, Vieira GC, Valiati VH, Deprá M, Valente VL da S. Combining morphology and molecular data to improve *Drosophila paulistorum* (Diptera, Drosophilidae) taxonomic status. Fly (Austin). 2018;12:81–94. https://doi.org/10.1080/19336934.2018.1429859.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 2717

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title “Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology”.)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-599019

ACTA UNIVERSITATIS
UPSALIENSIS
2026