Investigating relationships between sequence conservation and function using multispecies whole genome alignments

Feron Romain

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département d'Ecologie et Evolution

Investigating relationships between sequence conservation and function using multispecies whole genome alignments

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Romain FERON

Biologiste diplômé ou Master de l'Université de Montpellier 2

Jury

Prof. Nelly PITTELOUD, Présidente
Prof. Marc ROBINSON-RECHAVI, Directeur de thèse
Dr. Robert WATERHOUSE, Co-directeur de thèse
Prof. Laurent DURET, Expert
Dr. Johannes RAINER, Expert

Lausanne
2024

UNIL | Université de Lausanne
Faculté de biologie
et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Madame	Prof.	Nelly	**Pitteloud**
Directeur·trice de thèse	Monsieur	Prof.	Marc	**Robinson-Rechavi**
Co-directeur·trice	Monsieur	Dr	Robert	**Waterhouse**
Expert·e·s	Monsieur	Dr	Laurent	**Duret**
	Monsieur	Dr	Johannes	**Rainer**

le Conseil de Faculté autorise l'impression de la thèse de

Romain Feron

Master in Bioinformatics, Université de Rennes I, France

intitulée

Investigating relationships between sequence conservation and function using multispecies whole genome alignments

Lausanne, le 1 juillet 2024

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Nelly Pitteloud

Acknowledgements

The present thesis is the result of a long and complicated journey, both academically and personally. This journey has brought me many great and memorable moments, but also some of the most challenging times in my life. These challenges have helped me grow a lot both as a scientist and a person, but I would not have overcome them without support from a diversity of people to whom I now want to express my gratitude. It is without doubt to me that I would not have succeeded in completing this academic journey without them.

First and foremost, I want to express my appreciation for everything my supervisor Rob has done for me since I first contacted him in 2018 - and I am glad I did! From the very beginning of this PhD, Rob has made it clear I could open to him and count on his support whenever things were not going the way I wished they would, in and outside of work. Over the years, he offered me many opportunities to contribute to collaborative projects, and he supported all my own endeavours, particularly with teaching, going the extra mile to make sure I received credit for that work. Our scientific discussions are always stimulating and motivating, and I was able to work at my best under his supervision. I could not have asked more from my PhD advisor, and I certainly could have expected less; Rob, thank you for everything you did for me and your support throughout this experience!

It is thus only natural that the group Rob gathered over the years was a great environment, and I am grateful to all the group members who contributed to the positive and stimulating atmosphere in our office. I truly enjoyed spending time with you in and outside of work, and I'm looking forward to keeping doing so after the end of this chapter. I particularly want to thank Livio, who was for a long time the only other PhD student in the group; we shared most of this journey together, and we taught each other a lot. You often challenged my perspective on complex topics and helped me grow, and always pushed and supported me, which I really appreciate. I also want to thank Antonin for the many collaborations and exchanges we had, in particular on workflows and Snakemake.

It can be difficult to do your best work when experiencing challenging times, and with the timeline of this thesis being scarred with the COVID pandemic and the long period of social isolation it brought on people, the completion of this thesis manuscript is a perfect opportunity to express my gratitude to the people who made it possible to endure it all. There are too many to name them all, and I am thankful to all of them. However, I want to express my sincere appreciation to Christian, who has always been there for me when I needed him, has always told me what I needed to hear, even if I didn't want to, and often pushed me in the right direction. Similarly, I want to warmly thank Thibault, who has been a constant beacon of support for me over the years and gave me feedback on this manuscript. I also thank Edu and Giacomo who welcomed me in Villa Pina when I needed a place to stay, and who changed me for the better - physically and mentally; and a special thank you to Miya, to the early days of the lake aux rats, and to the petits suisses, for being my supportive friends in all these years. Finally, I am grateful for the constant love and support of my family, who was there when it mattered despite the distance.

As we all know, academic work is never solely the product of one’s individual efforts, and many people contributed to this project through collaboration, feedback, and discussions. I want to specifically thank Irwin Jungreis for his multiple insights on exploiting MWGA data and using PhyloCSF, as well as the NCBI datasets team for our interactions while building the A^3Cat. I am grateful to Corinne Dentan and Anne Cuendet for helping me navigate the administrative hardships of UNIL over the years, and to Marc for his administrative help with the thesis. Finally, I thank all members of my PhD committee, Nelly Pitteloud, Laurent Duret, Tanja Schwander, and Johannes Rainer for all their past and future feedback on this project.

Table of Contents

Summary

The main goal driving the work presented in this thesis is to investigate the relationship between sequence conservation and biological function in Arthropods. Indeed, the increasingly comprehensive sampling of all kingdoms of life enabled by developments in large-scale DNA sequencing and driven by large-scale sequencing initiatives brings powerful opportunities to explore patterns of genome evolution and characterise novel functional genomic elements using multi-species comparative genomics approaches. The basic premise of such approaches is that sequences that remain conserved or recognisably similar across many species over millions of years of evolution are constrained to do so because of evolutionary pressures to maintain some biologically functional role. Consequently, furthering our understanding of the relationship between these evolutionary constraints on genomic sequence and the biological function of the related genomic elements will facilitate the large-scale identification of functional elements in new genome assemblies, as well as generally strengthen our knowledge of how genomes encode biological function. A powerful method of detecting genomic sequences conserved across genomes relies on the computation of Multispecies Whole-Genome Alignments (MWGAs), which form the basic resource required to interrogate patterns of sequence changes and evolutionary constraints in relation to the functional spectra of genomic elements. Pioneering studies using signatures of evolutionary conservation to characterise functional elements first looked at the relatively small genomes of yeasts and *Drosophila*; following advances in sequencing technologies, further work investigated these patterns in mammals and angiosperms. However, despite the success of these studies, computing MWGAs remains a challenging task to this day, and at the start of this thesis project, there was no reliable implementation of the computational workflow required to do so. Furthermore, while the ever accelerating accumulation of available genome sequences enables increasingly powerful studies of evolutionary constraints on genomic sequence for more and more clades, the quality of these assemblies as well as the taxonomic coverage of sequence species remains heterogeneous even today.

In light of these observations, the first challenge addressed by this thesis work was to accurately assess both the quality and taxonomic distribution of genomic resources available for arthropods, in order to select genome assemblies to include in MWGAs. Our solution to this problem resulted in the release of an online resource powered by a computational workflow, the Arthropoda Assembly Assessment Catalogue (A^{3}Cat), which regroups all available information

on released and upcoming arthropod assemblies along with estimates of assembly quality computed by the workflow. This resource and the associated data was the focus of two publications, which are summarised in **Chapter 1**. The second task undertaken in this work was the development of bioinformatics tools and workflows necessary to build whole-genome alignment resources for arthropods as well as downstream analyses of sequence conservation and visualisation tools. In order to be useful to the general scientific community, these tools had to follow modern requirements for computational science by being portable, scalable, documented, and enabling fully reproducible computational analyses. This work resulted in two computational workflows implemented using a modern workflow management engine: one to compute MWGAs, presented in **Chapter 2**, and a second to perform analyses and generate powerful visualisation from MWGAs, described in **Chapter 3**; furthermore, additional work to develop specific missing blocks in the workflows and efforts to ensure reproducibility are covered in **Chapter 5**.

Because of the considerable challenges encountered during the development of the A^3Cat and the two computational workflows to generate and analyse MWGAs, amplified by the need for a reproducible implementation following modern practices, these technical advances comprise the main outcome of this thesis work. However, we were able to leverage this work to compute several MWGAs, which are described as part of **Chapter 2**, including a MWGA of 22 mosquito species which was used for a preliminary analysis of sequence conservation at the genome level presented in the results of **Chapter 3**. We expanded on the results of this analysis in **Chapter 4**, exploring how patterns of sequence conservation relate to biological function in protein-coding genes in mosquitoes using both long-term evolutionary conservation computed from the MWGA as well population-level genetic polymorphism. We focused particularly on genes whose products are involved in the immune system, as mosquitoes are vectors of some of the deadliest diseases to humans, identifying a link between specific functions in the immune system and sequence conservation for multiple gene families. Finally, we summarise additional work contributing to collaborative projects in multiple areas of genomics in **Chapter 6**.

Résumé

L'objectif principal des travaux présentés dans cette thèse est d'étudier le lien entre la conservation des séquences génomiques et la fonction biologique chez les arthropodes. En effet, l'échantillonnage de plus en plus complet des espèces vivantes, rendu possible par les progrès du séquençage de l'ADN et par les initiatives de séquençage à grande échelle, offre de puissantes opportunités d'explorer les schémas d'évolution du génome et de caractériser de nouveaux éléments génomiques fonctionnels à l'aide d'approches de génomique comparative multi-espèces. Le principe de base de ces approches est que les séquences qui restent conservées ou qui présentent des similitudes reconnaissables chez de nombreuses espèces au cours de millions d'années d'évolution sont contraintes de le faire en raison de pressions évolutives visant à maintenir un certain rôle biologiquement fonctionnel. Par conséquent, une meilleure compréhension de la relation entre ces contraintes évolutives sur la séquence génomique et la fonction biologique des éléments génomiques apparentés facilitera l'identification à grande échelle des éléments fonctionnels dans les nouveaux assemblages de génomes, et renforcera d'une manière générale notre connaissance de la manière dont les génomes codent pour une fonction biologique. Un outil puissant de détection des séquences génomiques conservées à travers les génomes repose sur l'élaboration d'alignements de génomes entiers multi-espèces (MWGA), qui constituent la ressource de base nécessaire pour interroger les modèles de changements de séquence et les contraintes évolutives en lien avec les catégories fonctionnelles d'éléments génomiques. Des études pionnières utilisant les signatures évolutives de la conservation de séquences pour caractériser les éléments fonctionnels ont d'abord porté sur les génomes relativement petits de la levure et de la drosophile ; d'autres travaux ont ensuite étudié ces modèles chez les mammifères et les angiospermes. Cependant, malgré le succès de ces études, la génération de MWGA reste à ce jour une tâche difficile, et au début de ce projet de thèse, il n'existait pas d'implémentation fiable du *workflow* nécessaire pour y parvenir. De plus, alors que l'accumulation toujours plus rapide de séquences génomiques permet des études de plus en plus étendues des contraintes évolutives sur le génome pour un nombre croissant de clades, la qualité de ces assemblages ainsi que la couverture taxonomique des espèces séquencées restent encore aujourd'hui hétérogènes.

À la lumière de ces observations, le premier défi relevé par ce travail de thèse a été d'évaluer avec précision la qualité et la distribution taxonomique des ressources génomiques disponibles

pour les arthropodes, afin de sélectionner les assemblages de génomes à inclure dans les MWGA. Notre solution à ce problème a abouti à la publication d'une ressource en ligne, l'*Arthropoda Assembly Assessment Catalogue* (A^3Cat), qui regroupe toutes les informations disponibles sur les assemblages d'arthropodes publiés et à venir, ainsi que les estimations de la qualité des assemblages calculées par le *workflow* générant cette ressource. Le catalogue et les données associées ont fait l'objet de deux publications, qui sont résumées dans le **Chapitre 1**. La deuxième tâche entreprise dans le cadre de ce travail a été le développement d'outils bioinformatiques et de *workflows* nécessaires pour créer des ressources d'alignement de génomes entiers d'arthropodes ainsi que des analyses en aval de la conservation des séquences et des outils de visualisation. Afin d'être utiles à la communauté scientifique, ces outils devaient répondre aux exigences modernes de la science informatique en étant portables, *scalables*, documentés et en permettant des analyses informatiques entièrement reproductibles. Ce travail a abouti à deux *workflows* mis en œuvre à l'aide d'un moteur moderne de gestion des *workflows* : l'un pour générer les MWGA, présenté dans le **Chapitre 2**, et l'autre pour effectuer des analyses et générer des visualisations à partir des MWGA, décrit dans le **Chapitre 3** ; en outre, des travaux supplémentaires visant à développer des blocs spécifiques manquants aux *workflows* et des efforts pour assurer la reproductibilité sont couverts dans le **Chapitre 5**.

En raison des défis considérables rencontrés lors du développement de l'A^3Cat et des deux *workflows* pour générer et analyser les MWGA, amplifiés par la nécessité d'une implémentation reproductible adhérant aux pratiques modernes, ces avancées techniques constituent le principal résultat de ce travail de thèse. Cependant, nous avons pu tirer parti de ce travail pour générer plusieurs MWGA, qui sont décrits dans le cadre du **Chapitre 2**, y compris un MWGA de 22 espèces de moustiques qui a été utilisé pour une analyse préliminaire de la conservation des séquences au niveau du génome, présentée dans les résultats du **Chapitre 3**. Nous avons développé les résultats de cette analyse au **Chapitre 4**, en explorant la manière dont les schémas de conservation des séquences sont liés à la fonction biologique des gènes chez les moustiques, en utilisant à la fois la conservation de séquence à l'échelle évolutive calculée à partir de ce MWGA et le polymorphisme génétique au niveau des populations. Nous nous sommes particulièrement intéressés aux gènes dont les produits sont impliqués dans le système immunitaire, car les moustiques sont les vecteurs de certaines des maladies les plus mortelles pour l'homme, et nous avons identifié un lien entre des fonctions spécifiques du système immunitaire et la conservation des séquences pour plusieurs familles de gènes. Enfin, nous résumons au **Chapitre 6** les travaux supplémentaires contribuant à des projets de collaboration dans de multiples domaines de la génomique.

Introduction

The work presented in this thesis details the development of computational tools and workflows designed to enable the comparative genomics community to fully exploit the rapidly growing amount of genome data. The thesis specifically describes the work performed to provide researchers with the means to efficiently and reproducibly build and analyse multispecies whole genome alignments (MWGAs), which allow comprehensive exploration and investigation of the relationships between conservation of genomic sequence and the biological function of genomic elements. In this introductory chapter, we first provide biological and technical context on functional genomic elements and the expectations for their conservation throughout evolution. We then present summaries of the theoretical frameworks underlying sequence alignment methods and the history of approaches developed to overcome the computational challenges of building whole genome alignments. Finally, we introduce community motivations driving the development of computational solutions to meet the increasingly recognised needs for open, accessible, and reproducible tools and workflows in modern science.

Functional genomic elements and how to find them

Genomes encode the building blocks of life

Understanding how the information encoded in genomes controls biological function is a major challenge in biology. In the early days of genomics, most of the work to address this question focused on identifying and characterising protein-coding genes, because these genomic elements have the most evident link to biological function and are the easiest to identify (Harrow et al., 2009). Nonetheless, protein-coding genes can have a complex structure: the entire gene sequence is transcribed into a messenger RNA but introns and untranslated regions (UTRs) are not translated into proteins. Furthermore, as the field of genomics progressed and whole genome sequences became available for several species, the importance of non-coding genomic elements in expressing and regulating biological functions came to light (Wright & Bruford, 2011). Many of these non-coding sequences, including promoters, enhancers, silencers, insulators, microRNAs, and long non-coding RNAs (lncRNAs) interact with transcription factors and other proteins in a complex spatial and temporal network to regulate the expression of genes (Spitz & Furlong, 2012). Some of these elements, for instance lncRNAs, can also have a direct role in specific cellular processes (Wilusz et al., 2009). To understand the intricacies of how genomes govern biological processes, it is therefore crucial to identify all functional sequences and the structural elements of which they are made (Maston et al., 2006). Following this idea, considerable efforts led by the Encyclopedia of DNA Elements (ENCODE) project were made to experimentally identify functional elements in the human genome and later in model organisms including *Caenorhabditis elegans* and *Drosophila melanogaster*, revealing that up to 80% of the human genome may be linked to a biological function (ENCODE Project Consortium, 2012). Such comprehensive efforts serve to define and refine the complete catalogue of functional genomic elements in a given genome, the building blocks that govern organismal biology.

Functional elements usually exhibit cross-species sequence conservation

Building comprehensive catalogues of functional genomic elements has been achieved for several well-studied species thanks to laborious and often expensive experimental work, however, this usually cannot be applied to most non-model species. One solution to overcoming this challenge was the development of computational methods to discover and annotate functional elements in genomes (Mathe, 2002). Some methods were tailored to specific classes of functional elements, but a successful general approach to finding functional elements consists of characterising patterns of conservation of genomic sequence across multiple species (Elgar & Vavouri, 2008). The hypothesis underlying this approach is that conserved sequences are under selective constraints acting on a biological function, and therefore locating these sequences is a major clue for the identification of functional elements in the genome of a species (Alföldi & Lindblad-Toh, 2013). This hypothesis has been at the basis of many approaches to identify specific types of functional sequences, and although not all functional sequences are conserved, it is generally accepted that most conserved sequences are functional (Pang et al., 2006). Conservation of genomic sequence has been studied in relation to model species in yeasts (Kellis et al., 2003), mammals (Lindblad-Toh et al., 2011), flies (Stark, Lin, et al., 2007), angiosperms (Hupalo & Kern, 2013), and in a few other non-model clades (Roux et al., 2014; X. Wang et al., 2015; Woodard et al., 2011). Studies in human and *Drosophila* in particular identified specific signatures of evolutionary conservation for different families of genomic elements and, in some cases, were able to link patterns of conservation to biological function, for instance highly conserved promoters being associated with developmental functions in humans (Woolfe et al., 2004). Early comparisons between multiple clades were performed using the limited available genomes at the time (Siepel et al., 2005), but overall, such insights have been lacking outside of these model species, and the lack of data on conservation of sequence in other species has prevented large-scale comparisons across multiple clades in an evolutionary framework. This scarcity is due at least in part to the taxonomic heterogeneity of available genomes and to the computational challenges involved in comparing whole genomes between multiple species.

Whole genome alignments enable identification of conserved sequences

Taxonomic sampling of genomes is improving in recent years, as technical advances have greatly decreased sequencing costs and improved assembly methods, leading to a rapidly increasing number of high-quality genome assemblies being generated for species across the tree of life. Ambitious initiatives like the Earth BioGenome Project (Lewin et al., 2018) aiming to sequence all identified species will further accelerate this trend in the coming years. The resulting abundance of high-quality assemblies is therefore filling the taxonomic gaps in available genomes and enabling the study of conservation of sequence in a multitude of taxa at different evolutionary timescales. Efforts to exploit these growing numbers of genomes in comparative analyses using Multispecies Whole Genome Alignments (MWGAs) are enabling genome-wide quantification of sequence conservation. For example, the Zoonomia consortium's whole-genome alignment of 240 placental mammals representing all orders was used to estimate that 10.7% of the human genome is evolutionarily conserved and to catalogue more than 4,500 ultraconserved elements (Christmas et al., 2023). Analysis of this placental mammal MWGA enabled the detection of 101 million exceptionally conserved (significantly constrained) single nucleotides in both coding and noncoding regions of the human genome that are thus likely to be functionally important. The formal identification of such conserved sequences using MWGAs proceeds by building two phylogenetic models, one for conserved regions based on sites expected to be conserved (e.g. the highly conserved first nucleotide of the three nucleotides that make up a codon in protein-coding sequences), and one for non-conserved sequences (e.g. the variable third nucleotide of the three nucleotides that make up a codon in protein-coding sequences). Each position in the MWGA (i.e. each column of nucleotides in the multiple alignment), can then be tested for whether it has a significantly better fit to the conserved or to the non-conserved phylogenetic model. The building and testing of such phylogenetic models (Siepel & Haussler, 2005) have been implemented in popular tools such as phastCons (Siepel et al., 2005) and PhyloP (Pollard et al., 2010). Results from using MWGAs to identify conserved sequences allow for the exploration of general signatures of evolutionary conservation for coding and non-coding functional genomic elements, which will facilitate the *de novo* annotation of these elements in existing and in future genome assemblies. When combined with functional genomics and other complementary data, opportunities arise to unravel the links between identified conservation signatures and putative biological processes, thereby helping to associate biological functions to these annotated conserved elements.

Theoretical background on sequence alignment methods

Sequence alignment constitutes a foundational process in many computational analyses of biological sequence data. The problem of optimal global alignment of two sequences was virtually solved in the 1970s and implemented for biological sequences by Needleman and Wunsch (Needleman & Wunsch, 1970); to better adapt to the reality of biological sequences, later algorithms like Smith-Waterman (Smith & Waterman, 1981) were designed to identify local alignments in a pair of sequences, rather than a unique best global alignment. However, these algorithms are computationally intensive and cannot easily be applied to long sequences or multiple pairs of sequences, let alone entire genomes; furthermore, aligning entire genomes requires handling duplications and rearrangements to reconstruct orthology relationships and not just homology. Consequently, alignments at the genome scale require heuristic and dedicated algorithms to find close-to-optimal alignments in a reasonable time, and such algorithms were developed to handle the escalating volume of biological data (Delcher, 2002; Delcher et al., 1999).

A popular approach is to identify approximate best local alignments using seeding-extension algorithms and then filter and process these local alignments to generate a whole-genome alignment. This process relies heavily on efficiently finding approximate best local alignments between two large collections of sequences, which is implemented in popular tools like the Basic Local Alignment Search Tool, BLAST (Altschul et al., 1990) and the BLAST-Like Alignment Tool, BLAT (Kent, 2002). However, these tools are optimised for closely-related sequences, and specific software like LASTZ (Harris, 2007) and LAST (Kiełbasa et al., 2011) were developed to align genomes of evolutionarily more distant species. LASTZ in particular has been used to compute pairwise alignments in most MWGAs available today, including the alignments used in conservation tracks from the University of California Santa Cruz (UCSC) Genome Browser. LASTZ was developed as a successor to the formerly-popular BLASTZ (Schwartz et al., 2000) to extend its functionalities, simplify parameter selection by the user, and optimise memory usage. The software implements a seeding strategy allowing base-pair transitions and user-specified seed patterns. Seed hits between two sequences are extended into ungapped *high-scoring segment pairs* (HSP) until the alignment score of the two sequences drops to a negative threshold. Long HSPs are merged into *chains* of consecutive overlapping HSPs and the resulting chains are integrated into gapped alignments using an anchoring process. The final output of LASTZ contains all extended gapped alignments between the two sequences found with this approach.

The aforementioned methodological developments apply to the alignment of two sequences, yet MWGAs require aligning multiple sequences, which is an NP-hard problem and therefore cannot be computed in a reasonable time for whole genomes without involving additional heuristics (L. Wang & Jiang, 1994). A common strategy to align multiple sequences is to perform progressive alignments using a tree describing the distance between sequences as a guide. Progressive alignments are well suited to multiple whole-genome alignments and are implemented in popular tools like progressiveMauve (Darling et al., 2010), Mugsy (Angiuoli & Salzberg, 2011), and MULTIZ (Blanchette et al., 2004), the latter being used to compute MWGAs for several seminal sequence conservation studies (Kellis et al., 2003; Lindblad-Toh et al., 2011; Stark, Lin, et al., 2007) and UCSC Genome Browser conservation tracks. MULTIZ takes as input pairwise alignments of each assembly to the assembly chosen as reference and uses a guide tree to progressively merge these alignments into a multiple genome alignment. One limitation of MULTIZ is that it can only produce reference-based alignments, which implies that an alignment between two non-reference sequences that is not alignable to the reference sequence will be discarded. This shortcoming spurred the development of new reference-free MWGA software, the main one being Cactus (Paten et al., 2011) which is still in active development. Cactus first computes local alignments with LASTZ and naively merges them into a graph-based multiple alignment; this alignment is aggressively filtered and non-aligned regions from the graph undergo a second, more sensitive alignment phase. The current version of Cactus is called progressiveCactus and uses a guide tree to partition the multiple sequence alignment problem in order to scale the algorithm to hundreds of genomes (Armstrong et al., 2020) and was used for the Zoonomia consortium's whole-genome alignment of 240 placental mammals (Christmas et al., 2023). Cactus performed well in early "Alignathon" methods assessments (Earl et al., 2014), is increasing in popularity, and will likely replace other MWGA approaches eventually, but it currently has several major limitations that make it not yet mature for large scale automated alignments. First and foremost, at the time of writing, Cactus is resource-intensive, and its usage and configuration are complex; it does not interface well with High Performance Computational platforms (HPCs) and its development is shifting towards cloud computing. Second, reference-free alignments are still not supported by most downstream software and need to be converted to reference-based multiple alignment format, which adds compute time and partly negates the advantages of Cactus in computing alignments.

History and current state of computing whole genome alignments

The basic premise of comparative genomics approaches is that sequences that remain conserved or recognisably similar across many species over millions of years of evolution are constrained to do so because of evolutionary pressures to maintain some biologically functional role (Alföldi & Lindblad-Toh, 2013). The very first pairwise and multi-species comparisons of eukaryote genomes demonstrated the power of this axiom, highlighting the utility – and methodological challenges – of whole-genome multiple sequence alignments for the discovery and characterisation of functional genomic elements (Ureta-Vidal et al., 2003). Pioneering studies targeted yeasts to take advantage of their small genomes, ~12 megabasepairs (Mb), where analysis of the genome alignments of several *Saccharomyces* species led to the revision of the yeast protein-coding gene catalogue and the identification of a suite of regulatory element motifs (Cliften et al., 2003; Kellis et al., 2003). With their much larger genomes, early studies on vertebrates focused on specific regions rather than whole genomes, e.g. developing phylogenetic shadowing techniques to analyse four regions from 13 to 17 primates enabled the discovery of primate-specific gene regulatory elements and the delineation of exons from multiple genes (Boffelli et al., 2003). Across 12 more evolutionarily diverse vertebrates, sequencing regions orthologous to a seven-gene-containing segment of about 1.8 Mb on human chromosome seven enabled the identification of multi-species conserved sequences, which showed conservation patterns indicating both functional constraints and neutral mutations, and included many novel conserved non-coding segments not evident from pairwise analyses (Margulies et al., 2003; Thomas et al., 2003).

The aim of the encyclopaedia of DNA elements (ENCODE) project “to identify all functional elements in the human genome” (ENCODE Project Consortium, 2004) undoubtedly helped to drive advances in sequence analysis approaches, e.g. by combining phylogenetic models of molecular evolution with hidden Markov models (Siepel & Haussler, 2005), and the development of robust approaches for the alignment of multiple whole genomes such as the threaded blockset aligner (TBA) (Blanchette et al., 2004). Employing these new tools to build and analyse four separate genome-wide multiple alignments identified evolutionary conserved elements in five vertebrates, four insects, two worms, and seven yeasts (Siepel et al., 2005). These elements covered 3%–8% of the human genome and much larger fractions of the more compact genomes of *Drosophila melanogaster* (37%–53%), *Caenorhabditis elegans* (18%–37%), and *Saccharomyces cerevisiae* (47%–68%). Focusing on 44 regions encompassing about 1% of the human genome, the ENCODE pilot project applied an array of experimental techniques to characterise human genome functions (The ENCODE Project Consortium, 2007). The functional genomics data were complemented by evolutionary analyses of multi-sequence alignments of the corresponding genomic regions from 23 other mammals to identify regions under evolutionary constraint, which spanned about 5% of the ENCODE nucleotides (Margulies et al., 2007). Employing several different alignment and constraint-measurement methods, these evolutionary analyses found ~40% of constrained sequence to be annotated protein-coding exons or their untranslated regions (UTRs), a further ~20% corresponded to other experimentally-identified functional elements, but the remaining ~40% was functionally unannotated.

Successful scaling up of the alignments to whole-genome levels required the development of computational pipelines such as those designed for building the alignment and conservation tracks at the University of California Santa Cruz (UCSC) Genome Browser, which now include the 240 Zoonomia consortium placental mammals (Nassar et al., 2023). Earlier analyses of their 28-way vertebrate whole-genome alignments highlighted the power of such approaches for exploring vertebrate genomic evolution, and showed how different types of elements (coding exons, regulatory regions, etc.) exhibited very different rates and modes in the decline of alignability at increasing phylogenetic distances (Miller et al., 2007). The advancement of multi-species analyses to examine patterns of conservation in increasing detail is exemplified by the study of 12 *Drosophila* genomes (Drosophila 12 Genomes Consortium, 2007). Applying evolutionary signature analyses to the 12-species genome alignment led to the *de novo* discovery of functional genomic elements with improved precision and sensitivity of evolutionary inferences (Stark, Lin, et al., 2007). This enabled the cataloguing of new and revised protein-coding genes and exons as well as numerous stop-codon readthrough events (Lin et al., 2007), the annotation of novel non-protein-coding genes such as microRNAs (miRNAs) (Stark, Kheradpour, et al., 2007), and the identification of miRNA target sites and other regulatory motifs (Kheradpour et al., 2007).

Further developments to these and other methodologies were applied to the alignment of 29 placental mammal genomes to chart a high-resolution map of human evolutionary constraint where constrained elements encompassed 4.2% of the genome (Lindblad-Toh et al., 2011). About 30% of these elements were associated with protein-coding transcripts (19.6% protein-coding sequences), ~27% overlapped mapped chromatin states (e.g. enhancers or insulators), ~3% regulatory motifs, and ~1.5% RNA structures, leaving just under 40% with no functional clues – in line with estimates from the ENCODE pilot project. In flowering plants, using the relatively compact genome of the thale cress *Arabidopsis thaliana* (119 Mb) as the reference onto which to build a 20-species whole-genome alignment helped to circumvent the challenges associated with duplications in polyploid plant genomes (Hupalo & Kern, 2013). Plant conserved elements showed about double the proportion of protein-coding regions (42%) compared with vertebrates, leaving 19% in intronic and 5% in other features (e.g. regulatory elements), and 33% in unannotated intergenic regions. In addition, comparing levels of conservation across the phylogeny in terms of the fractions of alignable base pairs revealed a slower decay of plant genome feature conservation than across vertebrates. A later study of 17 grass genomes identified at least 12% of the rice genome to be evolving under constraint, where comparisons with population polymorphism data showed that constrained sequences exhibited depleted single nucleotide polymorphism (SNP) frequencies (Liang et al., 2018).

The latest milestone in terms of taxonomic span and depth was achieved using progressiveCactus (Armstrong et al., 2020) to build the Zoonomia consortium's whole-genome alignment of 240 placental mammals (Christmas et al., 2023). The UCSC Cactus team formed a key part of the consortium and worked to compute the complete alignment set in HAL format, for the rest of the consortium to then use as the basis for their analyses. The analysis of the MWGA representing all orders of placental mammals showed that 10.7% (332 Mb) of the human genome is evolutionarily conserved and identified more than 4,500 ultraconserved elements (Christmas et al., 2023). The MWGA was used to measure constraint (significant conservation) across the human, chimpanzee, mouse, dog, and little brown bat reference genomes by projecting the Cactus alignment onto each species (HAL to MAF format) and then measuring sequence constraint with phyloP (Siepel et al., 2005). Constraint in the primate subset of the MWGA was assessed using phastCons (Hubisz et al., 2011), going beyond per-nucleotide constraint and allowing for the identification of conserved elements. A gene-focused analysis identified the most constrained genes as being enriched in functional processes such as post-transcriptional regulation of gene expression and embryonic development. In contrast, genes implicated in functions including innate and adaptive immunity, skin development, smell, and taste were amongst the most accelerated genes. These insights, and the many others reported by the Zoonomia consortium, into the relationships between sequence conservation and function exemplify the power of using MWGAs to investigate evolutionary constraints of functional genomic elements. However, the construction of large-scale MWGAs remains computationally extremely challenging, requiring in this case the direct support of the Cactus team to achieve the feat of aligning so many genomes. This was a primary motivation for the work of this thesis, to democratise the ability to build MWGAs by developing the necessary tools and workflows for assessing genome quality, processing myriad file format conversions and data preparation steps, introducing parallelisation where possible, and packaging downstream analysis tools into an ecosystem of informatics solutions for building and analysing MWGAs.

Reproducible tools and workflows are a requirement for modern science

There is a reproducibility and availability crisis in science

With the frantic pace required to keep up with the constant advances in our fields of knowledge while being subjected to unrelenting pressure to publish positive results, it is easy for one to lose track of the foundations of the process underlying science - or at least experimental science, to which biology belongs. This phenomenon has long been exacerbated by the requirements from funding bodies and publishing media, which overly emphasised the societal impact of results over technical soundness and adherence to the scientific method. While the exact model of this process has been debated and many variations exist (Nola & Sankey, 2014), an arguably general description of this process includes, in order, 1) observing existing data, 2) formulating a hypothesis, 3) designing and running an experiment testing this hypothesis, 4) analysing the data obtained from the experiment, 5) drawing conclusions from the analyses results, and 6) reporting these conclusions along with the entire process required to reach them. Because this process is complex, and to limit the impact of biases in experiment design and interpretation of results, it is essential that each step of this process undergoes careful scrutiny by knowledgeable and objective peers. In addition, the process should be replicated independently to provide support for - or conversely, to refute - the results and conclusions reached by a body of work. In practice, however, the overwhelming pressure to publish novel results lead to an almost complete absence of replicative studies (Makel et al., 2012); in fact, the volume of scientific publications makes even comprehensive peer validation effectively impossible in many cases. Perhaps in part because of this phenomenon, the number of scientific studies that could not be replicated - in cases where replication was attempted - has increased drastically (Begley & Ellis, 2012). Unreplicable studies fall within two categories: the first, and most serious one, results from unrigorous scientific practices, sometimes to the level of scientific fraud (Agnoli et al., 2017; Fraser et al., 2018; Simmons et al., 2011); the second, more widespread, comes from an insufficient transparency, level of details, and / or incomplete reporting of the methods used in the study (Bakker & Wicherts, 2011; Nuijten et al., 2016). Albeit with different severity, both categories contribute to the expansion of a body of unverified and unreplicable results in the scientific literature, leading to a general "reproducibility crisis" experienced by more than 90% of researchers (Baker, 2016). Interestingly, this crisis is well illustrated by the example of MWGAs, presented in the first part of this introduction: although several studies have computed and used MWGAs successfully as early as 2003, none of the associated work was directly replicated, and the methods used in these studies have seen very little use because of their complexity and lack of accessibility.

Reproducibility of computational analyses

Although the issue with reproducibility extends to the entirety of the scientific process, it is particularly present in the data analysis step. With the rapid development of computing since the 1990s, virtually every current scientific study in biology includes computer-based data analyses and visualisation, the complexity of which varies greatly, ranging from simple spreadsheet-based data tables to complex statistical analyses and visualisations involving extensive processing of large-scale data. By their nature, and in contrast with lab-based or field-based scientific work, computer-based analyses are in a unique advantageous position with regards to reproducibility. Indeed, both raw and processed data can be shared infinitely at moderate cost, and analyses can be automated due to their reliance on programming and scripting. Consequently, initiatives to develop software and guidelines for the reproducibility of data analyses started relatively early (Hoon et al., 2003; Oinn et al., 2004), leading to a definition of criteria for reproducibility: to be reproducible, an analysis should be *automatable* with minimal efforts from a user attempting to replicate it; it should be *portable* and *scalable*, making it executable on - ideally - any platform and system; it should be *transparent* and *documented*, exposing the entire process involved in the analysis, including software, parameters, and code; and finally, the data used in the analysis should be *accessible*. Building on these criteria, comprehensive systems evolved into multiple frameworks handling every aspect of reproducibility for computational analyses. Among these systems, the most popular today are Nextflow (Di Tommaso et al., 2017), Snakemake (Köster & Rahmann, 2012), Galaxy (The Galaxy Community et al., 2022), with the meta-language Common Workflow Language attempting to facilitate compatibility between systems (Crusoe et al., 2022). These solutions attempt to resolve the issue with reproducibility of computational analyses with a different approach, each with its strengths and weaknesses, and overall they greatly contributed to the improvements in this direction. In parallel to the development of these systems, a second approach focused specifically on *portability*, which is arguably the most difficult problem to solve in reproducibility. Indeed, many tools used in computational analyses are only available for specific systems or configurations, their source code is not open, and they sometimes simply do not provide an easy installation solution (Aron et al., 2021). Even when they are available, changes in systems and dependencies, as well as updates to software, can greatly hinder the reproducibility of analyses (Kern et al., 2020). Multiple systems were designed to specifically address this problem, two of the most popular in scientific computing are Docker (Merkel, 2014), a general container solution, and Bioconda, a Conda channel managing packages for bioinformatics software (Grüning et al., 2018). Both Conda and Docker were integrated in the major frameworks Nextflow and Snakemake and are the main solution to resolving *portability* in these two systems.

Managing and organising the rapid accumulation of available genomic data

The problem of reproducibility of computational analyses is exacerbated by the accelerating pace at which data accumulates; this particularly applies to large-scale genomic data, including high quality genome assemblies for an increasingly wide sampling of the tree of life, individual resequencing data, gene expression data, and many other resources. Robust open and reproducible computational workflows will play a key role in leveraging the power of these data through large scale comparative genomic analyses; without rigorous efforts to maintain discipline in the reproducibility of analyses, we are bound to accumulate unverified, potentially erroneous results building upon each other (Cohen-Boulakia et al., 2017). Yet, managing this avalanche of data requires more than reproducible workflows: existing and yet to be produced resources need to be openly organised, managed, annotated, and made available to the entire scientific community. This idea was formalised into a framework through the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles (Wilkinson et al., 2016), which are used by large initiatives and institutes generating and managing genomic data such as the Earth BioGenome Project (EBP) (Lewin et al., 2018), the Darwin Tree of Life project (The

Darwin Tree of Life Project Consortium, 2022), the NCBI GenBank (Sayers et al., 2019), or the European Nucleotide Archive (Leinonen et al., 2011). However, while these consortia and institutes have achieved much in managing and making accessible both primary and secondary genomic data, there is still a need for higher-level resources to improve the findability and organisation of these data. In the context of this thesis project, for instance, we often needed to survey available genome assemblies for arthropods to identify clades for which a sufficient number of high-quality assemblies were available to compute MWGAs. The information needed for this survey was mostly available in the form of metadata accessible on the NCBI, but some information, for instance on assembly quality, was only released as non-accessible data attached to publications. This observation led us to develop our own resource, the Arthropoda Assembly Assessment Catalogue (A^{3}Cat), which provides and organises comprehensive information on existing and upcoming arthropod genome assemblies in a searchable and filterable framework. Initiatives like the A^{3}Cat will be crucial to improve the *findability* and usability of genomic resources and will hopefully be integrated into existing resources in the future.

Summary of the thesis structure

The bioinformatics research work described in this thesis is presented in six chapters following this global introductory chapter. In **Chapter 1**, the background and motivations behind the development of the Arthropoda Assembly Assessment Catalogue (A^{3}Cat) resource are presented, with a summary of the approaches taken and the results achieved, as well as how this supports the rest of the thesis work and the broader research community. A major accomplishment of the thesis work is presented in **Chapter 2**: a comprehensive technical description of the development of the bioinformatics workflow required for building multispecies whole genome alignments (MWGAs), along with the underlying rationale and demonstrations of the application of the workflow to arthropod genome datasets. **Chapter 3** then goes on to describe the details of a complementary workflow that enables the downstream processing of MWGAs to compute key metrics through the integration of third-party software for analysing evolutionary constraint and protein coding potential, as well as tools that facilitate the visualisation of the results. In **Chapter 4**, the utility of the workflows and tools developed and described in the preceding chapters are demonstrated by applying them to the preliminary exploration of relationships between gene conservation and gene function using the mosquito immune system as a case study. Many challenges faced while developing the workflows described in chapters 2 and 3 required specific technical solutions which are detailed in **Chapter 5**, including a suite of tools developed to efficiently process MWGAs, a phylogenomics species tree reconstruction workflow, solutions for genome data track visualisations, as well as software packaging efforts and solutions for efficient use of high-performance computing facilities. Finally, **Chapter 6** summarises contributions made to complementary research projects undertaken during the course of the thesis including the development of tools, visualisations, and workflows to study the genetic mechanisms of sex determination as well as several projects investigating arthropod evolutionary genomics. The thesis concludes with a brief overview of the challenges addressed and how the methodological and technical solutions contribute to the advancement of the field in general, as well as the consequent future perspectives raised in the context of reproducible science.

Chapter 1: Surveying the current landscape of arthropod assemblies with the Arthropoda Assembly Assessment Catalogue

Summary

This thesis chapter summarises the research and development work performed to build a comprehensive quality assessment workflow, applied to publicly available genome assemblies of arthropods, with the results being made accessible through a web-browsable resource. The technical details are described in a research article published at GigaScience: "*Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes*" (Feron & Waterhouse, 2022a), included as **Appendix 1** of this thesis. The results of the work were also used as a basis for a review article published at Current Opinions in Insect Science: "*Exploring new genomic territories with emerging model insects*" (Feron & Waterhouse, 2022b), included as **Appendix 2** of this thesis. The outputs of this chapter provide the community with valuable resources: (i) a reproducible workflow that others can use to build their own assembly assessment catalogues, and (ii) a public online catalogue of comprehensively and consistently assessed arthropod genome assemblies. The results also serve as a foundation for the work presented in **Chapter 2**, allowing for the selection of only high-quality genomes for inclusion in multispecies whole genome alignment datasets.

Background and Motivation

Technical advances in sequencing technologies bringing down costs and reducing input sample requirements have led to an accelerating accumulation of new and improved genome assemblies (Hotaling et al., 2021). Efforts led by ambitious initiatives to sequence all known species, for instance the Earth BioGenome Project (EBP) (Lewin et al., 2018) and Darwin Tree of Life project (The Darwin Tree of Life Project Consortium, 2022), mean that this accumulation is likely to increase in the future. Arthropods display a remarkable phenotypic diversity which makes them attractive to study a wide range of topics including sociality, ageing, or ecosystem dynamics, and they can be agricultural pests or vectors of diseases. Because they are so diverse and studied for many research questions, arthropods will likely constitute a large proportion of upcoming genome assemblies, and the sequencing of arthropod species is supported by specific initiatives like the Global Invertebrate Genomics Alliance (GIGA Community of Scientists, 2014), the Global Ant Genomics Alliance (Boomsma et al., 2017), Arthropod genomics 100 pest genomes initiative (Childers et al., 2021), or the 5,000 insect genomes initiative (i5K Consortium, 2013), operating under the umbrella of the Earth BioGenome Project (Lewin et al., 2018).

One of the principal aims of this thesis project was to build the necessary computational workflows to generate MWGAs for multiple arthropod clades; high quality assemblies are key to compute MWGAs and to estimate conservation of sequence across species. In order to select which assemblies from which species to include in our MWGAs, we need a simple, comprehensive, and updatable assessment of the taxonomic coverage and quality of available arthropod genome assemblies. However, assessing the quality of an assembly is challenging because assemblies are models of an unknown truth. A first approach consists of computing metrics summarising the continuity of the assembly, *e.g.* N50 (length of the shortest contig so that 50% of the assembly is included in contigs longer than this value) and L50 (smallest number of contigs that make up 50% of the assembly), using software like Quality Assessment Tool for Genome Assemblies (QUAST, Gurevich et al., 2013). A complementary approach focuses on assessing assembly completeness by looking at gene or protein content using tools like Dual Organellar GenoMe Annotator (DOGMA, Dohmen et al., 2016; Wyman et al., 2004), Core Eukaryotic Genes Mapping Approach (CEGMA, Parra et al., 2007), or Benchmarking Universal Single-Copy Orthologues (BUSCO, Simão et al., 2015; Waterhouse et al., 2019). Among these, BUSCO has emerged as the standard and is now used by UniProt (*Assessing Proteome Completeness and Quality*, *n.d.)* and by quality assessment pipelines like BlobToolKit (Challis et al., 2020). Briefly, BUSCO relies on the idea that some genes are present in a single copy in almost all species within a lineage; such genes are called Universal Single-Copy Orthologues (USCOs), and sets of USCOs are manually curated for multiple lineages based on orthology data from OrthoDB (Waterhouse et al., 2011). BUSCO uses the software MetaEuk (*MetaEuk—Sensitive, High-Throughput Gene Discovery, and Annotation for Large-Scale Eukaryotic Metagenomics* | *Microbiome*, n.d.)- or optionally, Augustus (Stanke et al., 2006) - to predict all genes from a dataset in an assembly and output a score that represents the proportion of these genes that was successfully identified. The higher this proportion, the more likely the geneset of the evaluated assembly is complete.

Approach and Results

Most assemblies released in recent years use both continuity metrics and BUSCO scores as measures of quality, but these metrics are not yet available together in an easily accessible resource; furthermore, the computation of BUSCO scores can vary between assemblies when different versions, parameters, or lineage datasets are used. To address these shortcomings and generate a comprehensive, standardised quality assessment for all arthropod assemblies, we have developed an automated workflow applied for all arthropod assemblies available in the United States National Center for Biotechnology Information (NCBI) GenBank database. The workflow concurrently collates available NCBI assembly statistics and metadata and assesses assembly completeness with BUSCO to build a comprehensive catalogue of metrics in a taxonomically-aware framework. It is designed to be run regularly in order to update the collated resources to reflect new additions or changes at NCBI. We applied our workflow to assess 2,083 NCBI arthropod genome assemblies representing 1,387 species and to build the first release of the Arthropod Assembly Assessment Catalogue (A^3Cat) which we used to survey the current taxonomic coverage and assembly quality across arthropods. The workflows implemented to build the A^3Cat and the analyses of the first release are described in the first annexed publication (Feron & Waterhouse, 2022a, **Appendix 1**); we also used these data to review the current state of arthropod assemblies in the second annexed review publication (Feron & Waterhouse, 2022b, **Appendix 2**).

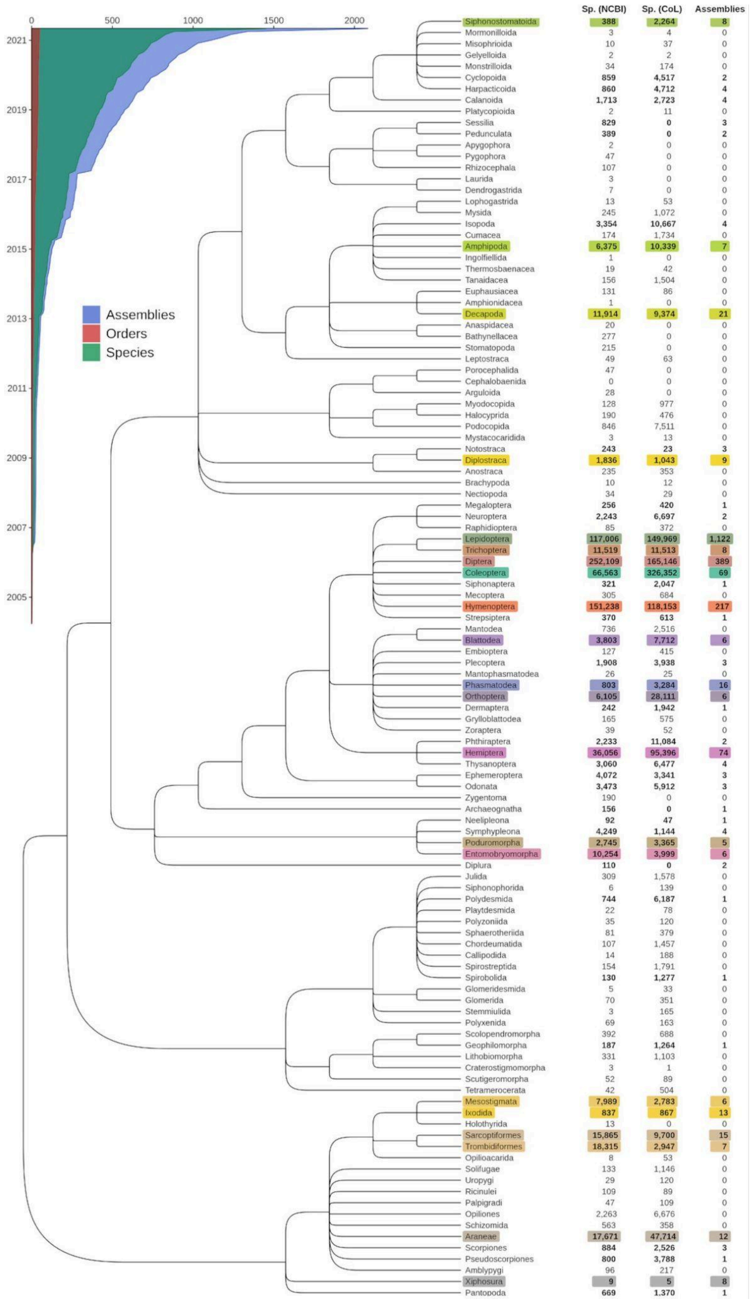
The first release of A^3Cat, including the aforementioned 2,083 assemblies, was published online on June 11th, 2021. Since then we updated the catalogue regularly; changes to the workflow, notably updating BUSCO from version 4 to the newer version 5.4.0, delayed some of these updates, but we are now attempting to maintain a monthly update schedule. Each release updates the main table with new assemblies submitted to NCBI and the figures showing the distribution of assembly metrics between arthropod orders. All BUSCO results as well as archived tables and JSON files from previous A^3Cat releases are archived and available for download. The publication presents a description of our automated analysis workflow that surveys genome assemblies, assesses their completeness using BUSCO, and collates the results into an interactively browsable resource (Feron & Waterhouse, 2022a). Using these results, we surveyed current taxonomic coverage and assembly quality at the NCBI. This survey highlighted the sparsity and taxonomic imbalance of current species sampling, with 79.5% of species (83% of assemblies) belonging to only 3 orders: Lepidoptera—e.g., butterflies, moths (712 species, 1,122 assemblies), Diptera—e.g., flies (216 species, 389 assemblies), and

Hymenoptera—e.g., ants, bees, wasps (175 species, 217 assemblies). Roughly half (142) of species with multiple assemblies were represented by a chromosome-level assembly, and across all assemblies, those labelled as chromosome-level accounted for 12.3%, while a further 41.1% were labelled as scaffold-level assemblies, and the remaining 46.6% were at contig-level. We also examined how key assembly metrics relate to gene content completeness. The EBP criteria for a reference-quality assembly include obtaining a complete and single-copy BUSCO score >90% and having the majority of sequences assigned to chromosomes. While 828 of the assessed arthropod assemblies achieved a complete and single-copy BUSCO score >90%, only 229 of these were also labelled as chromosome-level assemblies. Indeed, comparing assembly N50 values with their completeness scores showed that obtaining >90% complete BUSCOs can be achieved across a wide range of contiguities. While some with N50s <10 kb were able to achieve >90% or 80–90% completeness, the vast majority of assemblies with such low contiguity levels achieved considerably lower BUSCO completeness scores than more contiguous assemblies. The smallest assembly with a >80% Arthropoda completeness score was that of a grasshopper, however, inspecting the metadata revealed this to be a transcriptome and not a genome assembly.

The large number of assemblies included in our A^{3}Cat release accompanying the publication allowed us to comprehensively compare results from using different BUSCO lineage datasets and different BUSCO versions. Comparing percentages of complete BUSCOs identified with the Eukaryota (n=255) and the Arthropoda (n=1'013) lineage datasets for a total of 1'977 arthropod assemblies showed highly linearly correlated scores, especially for the highest-scoring assemblies. For those scoring <80% there was a small but noticeable shift towards Arthropoda producing slightly higher scores than Eukaryota, indicating that proportionately more of the larger set of Arthropoda BUSCOs can be recovered from lower-quality assemblies. Outlier points above the identity (y=x) axis (**Figure 4.A** in the attached publication 1) suggest that the lower-resolution Eukaryota lineage dataset occasionally produces overestimates of completeness, where proportionately more of the smaller set of ancient Eukaryota BUSCOs are recovered. Similar trends were observed when comparing the Arthropoda results to the higher-resolution Insecta (n=1,367) lineage dataset, with highly linearly correlated scores and occasional small overestimates of completeness using the Arthropoda lineage dataset. Comparing Arthropoda results to those from insect order-level lineage datasets revealed some shifts in completeness estimations that likely arise from the uneven representations of these orders in the 90-species Arthropoda lineage dataset, which is dominated by 20 hymenopterans and 15 dipterans. Using our large dataset to compare BUSCO v4 estimates with the previous BUSCO v3 results also highlighted interesting trends, revealing high levels of agreement for the highest-scoring assemblies, but a consistent shift towards lower scores reported by BUSCO v4 for lower-quality assemblies. Our large-scale analyses showed that the different BUSCO versions produce generally consistent estimates of completeness, with a tendency for the OrthoDB-v10–based Arthropoda and Insecta datasets to report lower scores, especially for lower-quality assemblies. For objective quantitative comparisons, it is therefore critical to assess assemblies using the same BUSCO versions, parameters, and lineage datasets, as presented in our A^{3}Cat resource for phylum-wide assessments of available arthropod genome assemblies. Finally, since the publication of A^{3}Cat, the Wellcome Sanger Institute has developed a search engine for genomic and sequencing project metadata across the eukaryotic tree of life called Genomes on a Tree (Challis et al., 2023) Application Programming Interface (API). To provide the arthropod genomics community with not only a catalogue of available assemblies but also an overview of arthropod species with ongoing and planned genome projects, we implemented an additional page in our A^{3}Cat resource displaying information on upcoming assemblies using a query to GoaT. Thanks to this update, the A^{3}Cat now presents exhaustive metadata and quality metrics for publicly available arthropod genome assemblies and summarises current efforts to target new species and accumulate high quality genomics data, making it a comprehensive hub for anyone interested in arthropod assemblies. Over the course of the project, we updated the A^{3}Cat 22 times, and it is still updated today (**Table 1.1**); the latest release of the A^{3}Cat (March 2024) now provides data quality and metadata for 5,792 assemblies from 3,412 species (**Figure 1.1**), as well as 4,308 species with ongoing or planned genome generation efforts.

Date	BUSCO version	Assemblies
2021-06-11	4.1.4	2,083
2022-01-11	4.1.4	2,699
2022-01-12	5.4.0	2,699
2022-04-27	5.4.0	2,987
2022-05-25	5.4.0	3,125
2022-07-30	5.4.0	3,298
2022-12-02	5.4.0	3,514
2023-02-01	5.4.0	3,826
2023-03-01	5.4.0	3,900
2023-04-01	5.4.0	3,974
2023-05-01	5.4.0	4,149
2023-06-02	5.4.0	4,381
2023-07-01	5.4.0	4,504
2023-08-01	5.4.0	4,696
2023-09-05	5.4.0	4,803
2023-09-21	5.4.0	4,830
2023-10-05	5.4.0	4,853
2023-11-06	5.4.0	4,951
2023-12-01	5.4.0	5,064
2024-01-02	5.4.0	5,199
2024-02-01	5.4.0	5,627
2024-03-01	5.4.0	5,792

Table 1.1: date, BUSCO version, and number of assemblies included in each update of the A^{3}Cat. Starting from February 2023, development was mostly done and the update process was streamlined, leading to regular monthly updates to this day. The number of Arthropod assemblies assessed in the catalogue - and thus deposited to NCBI Genbank - almost tripled since the start of the project. The workflow and website were originally developed using the widely-used version 4.1.4 of BUSCO; after version 5 was released, we waited for development to stabilise and public adoption to increase before settling on version 5.4.0 to update A^{3}Cat.



	Sp. (NCBI)	Sp. (CoL)	Assemblies
Siphonostomatoida	388	2,264	8
Mormonilloida	3	4	0
Misophrioida	10	37	0
Gelyelloida	2	2	0
Monstrilloida	34	174	0
Cyclopoida	859	4,517	2
Harpacticoida	860	4,712	4
Calanoida	1,713	2,723	4
Platycopioida	2	11	0
Sessilia	829	0	3
Pedunculata	389	0	2
Apygophora	2	0	0
Pygophora	47	0	0
Rhizocephala	107	0	0
Laurida	3	0	0
Dendrogastrida	7	0	0
Lophogastrida	13	53	0
Mysida	245	1,072	0
Isopoda	3,354	10,667	4
Cumacea	174	1,734	0
Amphipoda	6,375	10,339	7
Ingolfiellida	1	0	0
Thermosbaenacea	19	42	0
Tanaidacea	156	1,504	0
Euphausiacea	131	86	0
Amphionidacea	1	0	0
Decapoda	11,914	9,374	21
Anaspidacea	20	0	0
Bathynellacea	277	0	0
Stomatopoda	215	0	0
Leptostraca	49	63	0
Porocephalida	47	0	0
Cephalobaenida	0	0	0
Arguloida	28	0	0
Myodocopida	128	977	0
Halocyprida	190	476	0
Podocopida	846	7,511	0
Mystacocaridida	3	13	0
Notostraca	243	23	3
Diplostraca	1,836	1,043	9
Anostraca	235	353	0
Brachypoda	10	12	0
Nectiopoda	34	29	0
Megaloptera	256	420	1
Neuroptera	2,243	6,697	2
Raphidioptera	85	372	0
Lepidoptera	117,006	149,969	1,122
Trichoptera	11,519	11,513	8
Diptera	252,109	165,146	389
Coleoptera	66,563	326,352	69
Siphonaptera	321	2,047	1
Mecoptera	305	684	0
Hymenoptera	151,238	118,153	217
Strepsiptera	370	613	1
Mantodea	736	2,516	0
Blattodea	3,803	7,712	6
Embioptera	127	415	0
Plecoptera	1,908	3,938	3
Mantophasmatodea	26	25	0
Phasmatodea	803	3,284	16
Orthoptera	6,105	28,111	6
Dermaptera	242	1,942	1
Grylloblattodea	165	575	0
Zoraptera	39	52	0
Phthiraptera	2,233	11,084	2
Hemiptera	36,056	95,396	74
Thysanoptera	3,060	6,477	4
Ephemeroptera	4,072	3,341	3
Odonata	3,473	5,912	3
Zygentoma	190	0	0
Archaeognatha	156	0	1
Neelipleona	92	47	1
Symphypleona	4,249	1,144	4
Poduromorpha	2,745	3,365	5
Entomobryomorpha	10,254	3,999	6
Diplura	110	0	2
Julida	309	1,578	0
Siphonophorida	6	139	0
Polydesmida	744	6,187	1
Playtdesmida	22	78	0
Polyzoniida	35	120	0
Sphaerotheriida	81	379	0
Chordeumatida	107	1,457	0
Callipodida	14	188	0
Spirostreptida	154	1,791	0
Spirobolida	130	1,277	1
Glomeridesmida	5	33	0
Glomerida	70	351	0
Stemmiulida	3	165	0
Polyxenida	69	163	0
Scolopendromorpha	392	688	0
Geophilomorpha	187	1,264	1
Lithobiomorpha	331	1,103	0
Craterostigmomorpha	3	1	0
Scutigeromorpha	52	89	0
Tetramerocerata	42	504	0
Mesostigmata	7,989	2,783	6
Ixodida	837	867	13
Holothyrida	13	0	0
Sarcoptiformes	15,865	9,700	15
Trombidiformes	18,315	2,947	7
Opilioacarida	8	53	0
Solifugae	133	1,146	0
Uropygi	29	120	0
Ricinulei	109	89	0
Palpigradi	47	109	0
Opiliones	2,263	6,676	0
Schizomida	563	358	0
Araneae	17,671	47,714	12
Scorpiones	884	2,526	3
Pseudoscorpiones	800	3,788	1
Amblypygi	96	217	0
Xiphosura	9	5	8
Pantopoda	669	1,370	1

Figure 1.1. Taxonomic coverage of the Arthropod Assembly Assessment Catalogue A^{3}Cat v.2023-03-01. The Arthropoda phylogeny from the US NCBI Taxonomy database shows the evolutionary relationships amongst 114 orders. Counts of described species (Sp.) within each order are shown from the NCBI (v.2021–06-11) and the Catalogue of Life (CoL, v.2021–06-10), alongside numbers of genome assemblies available from the NCBI Assembly database (accessed on 25 August 2021). Of the 114 orders recognized by both the NCBI and the CoL, 48 orders are represented by ≥1 genome assembly. The 21 orders with ≥5 assemblies are highlighted with distinct colours.

Our development of A^{3}Cat led to an invitation to contribute to a special issue on Insect Genomics for Current Opinion in Insect Science. This special issue focused on how accumulating genomics resources are facilitating a shift from traditional model organisms to new model species or groups of model species for studying a large variety of biological phenomena at many different levels. In our review, we use results from A^{3}Cat to highlight how new genome resources are supporting emerging model systems that are advancing our understanding of insect biology and evolution. We show that while the quality of genome assemblies varies in contiguity and gene content completeness, technological advances are generally supporting new models by delivering high-quality genomic data. These new reference genomes themselves provide the framework onto which new knowledge can be mapped, from comparative genomic analyses, molecular biology experiments, as well as functional and population genomic datasets - transcriptomics, proteomics, metabolomics, resequencing, etc. Building on this idea, we used the number of NCBI BioProjects as a proxy to gauge the extent of genome-enabled research activities, and found that the classical model insect species, *Drosophila melanogaster*, is associated with an order of magnitude more registered projects than the other most represented species. Among the others are well-known species that are economically important, vectors of human diseases, or agricultural pests, all of which have had publicly available draft assemblies for more than five years and almost all of which now have published high-quality assembly upgrades, including most recently for the fall armyworm (Zhang et al., 2020), the tiger mosquito (Palatini et al., 2020), the brown planthopper (Ye et al., 2021), and the red flour beetle (Herndon et al., 2020). Species representing emerging model systems we highlighted in the review are expected to similarly build genome-anchored knowledge bases that support and enrich the exploration of insect diversity. Based on these observations, we concluded that gene content completeness and other quality assessments during production and of the resulting chromosome-level assemblies will continue to play a key role in establishing genome resources that best support the development of new model systems and advance understanding of insect biology and evolution.

Conclusion and Perspectives

The workflow and the resource produced as a result of this work represent important contributions to the field. The careful design and implementation of the workflow mean that it is easily deployable by others, as evidenced by the continued use and maintenance of the A^{3}Cat resource by the Waterhouse group in the context of the SNSF Sinergia project on arthropod moulting. The online A^{3}Cat resource is popular amongst arthropod genomics researchers, with a total of 2,151 page views from 689 unique users in the year 2023, and many personal communications from appreciative users around the world. A recent update to the BUSCO assessment tool included the ability to use a different, faster, method for the gene-finding step. Therefore future development of the workflow would need to incorporate minor changes to the workflow in order to take advantage of this new feature that promises to speed up the assessment process. Importantly, as well as serving the community the results presented in this chapter serve as an important basis for additional work performed as part of this thesis, as described in **Chapter 2** (for multispecies whole genome alignments) and **Chapter 5** (for species tree reconstructions).

Appendix 1: Research publication in GigaScience

Publication

Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes

Feron and Waterhouse. Gigascience. 2022 Feb 25;11:giac006. doi:10.1093/gigascience/giac006. PMID: 35217859

Author contribution

I developed the workflows, performed the analyses, including the major update to reflect BUSCO updates and the subsequent version comparisons, produced the figures, implemented the online resources, and contributed to writing and editing the manuscript.

OXFORD (GIGA)^n SCIENCE

GigaScience, 2022, **11**, 1–10
DOI: 10.1093/gigascience/giac006
TECH NOTE

Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes

Romain Feron [1,2] and Robert M. Waterhouse [1,2,*]

[1]Department of Ecology and Evolution, Le Biophore UNIL-Sorge, University of Lausanne, Lausanne 1015, Switzerland
[2]Evolutionary-Functional Genomics Group, L'Amphipole UNIL-Sorge, Swiss Institute of Bioinformatics, Lausanne 1015, Switzerland
*Correspondence address. Robert M. Waterhouse, Department of Ecology and Evolution, Le Biophore UNIL-Sorge, University of Lausanne, Lausanne 1015, Switzerland. E-mail: robert.waterhouse@unil.ch

Abstract

Background: Ambitious initiatives to coordinate genome sequencing of Earth's biodiversity mean that the accumulation of genomic data is growing rapidly. In addition to cataloguing biodiversity, these data provide the basis for understanding biological function and evolution. Accurate and complete genome assemblies offer a comprehensive and reliable foundation upon which to advance our understanding of organismal biology at genetic, species, and ecosystem levels. However, ever-changing sequencing technologies and analysis methods mean that available data are often heterogeneous in quality. To guide forthcoming genome generation efforts and promote efficient prioritization of resources, it is thus essential to define and monitor taxonomic coverage and quality of the data.

Findings: Here we present an automated analysis workflow that surveys genome assemblies from the United States NCBI, assesses their completeness using the relevant BUSCO datasets, and collates the results into an interactively browsable resource. We apply our workflow to produce a community resource of available assemblies from the phylum Arthropoda, the Arthropoda Assembly Assessment Catalogue. Using this resource, we survey current taxonomic coverage and assembly quality at the NCBI, examine how key assembly metrics relate to gene content completeness, and compare results from using different BUSCO lineage datasets.

Conclusions: These results demonstrate how the workflow can be used to build a community resource that enables large-scale assessments to survey species coverage and data quality of available genome assemblies, and to guide prioritizations for ongoing and future sampling, sequencing, and genome generation initiatives.

Keywords: arthropod genomes, biodiversity genomics, BUSCO assessments, genome assembly, genome quality database, reproducible workflow

Introduction

Advances in sequencing technologies are bringing down costs and reducing sample requirements, leading to an accelerating accumulation of new and improved genome assemblies. Ambitious initiatives to coordinate sequencing of all known species are generating representative genomes from across the tree of life that catalogue Earth's genetic biodiversity. In addition to constituting an inventory of biological diversity, the assembled and annotated genomes drive research to understand function and evolution at multiple levels, as well as to benefit human welfare [1, 2]. Investigating such questions using genomic data often requires comprehensive multi-species comparative analyses that benefit from high-quality assemblies [3, 4]. It is therefore essential to be able to define the current taxonomic coverage of high-quality assemblies to guide forthcoming sequencing efforts and promote efficient prioritization of resources globally.

Methods to gauge assembly quality include 2 main families of metrics [5]. One summarizes contiguity using metrics like N50 length, where half the assembly comprises sequences of length N50 or longer, or L50 count, the smallest number of sequences whose lengths sum to 50% of the assembly. Complementary approaches estimate completeness by examining gene or protein content, e.g., the DOmain-based General Measure for transcriptome and proteome quality Assessment (DOGMA) [6, 7] or BUSCO [8, 9]. BUSCO has emerged as a standard and is used by UniProt [10] and the US NCBI [11], as well as by genomics data quality assessment pipelines like MultiQC [12] and BlobToolKit [13]. BUSCO is based on the evolutionary expectation that single-copy orthologues found in nearly all species from a given taxon should be present and single-copy in any newly sequenced species from the same clade. BUSCO datasets are built for multiple taxonomic lineages by identifying near-universal groups of single-copy orthologues from OrthoDB [14, 15]. For assembly evaluations, sequence searches followed by gene predictions and orthology classifications identify complete, duplicated, or fragmented BUSCOs. The proportions recovered indicate the completeness in terms of expected subsets of evolutionarily conserved genes. Extrapolating from these, a high BUSCO completeness score suggests that the sequencing and assembly procedure has successfully reconstructed a reliable representation of the full set of genes.

Using their Complete Proteome Detector algorithm, UniProt classifies proteomes as "standard," "close to standard," or "outlier" and provides BUSCO proteome completeness summaries. For assemblies, the NCBI Assembly database provides summary statistics and metadata for each record. Querying these can provide snapshots of taxonomic coverage and data quality, but researchers currently lack access to comprehensive and standardized assessments of available assemblies. These would allow data producers to compare their assemblies to existing data at the most relevant taxonomic level. They would also provide re-

Received: October 27, 2021. **Revised:** December 12, 2021. **Accepted:** January 13, 2022

searchers with comprehensive overviews of resources for their focal taxa. Such communities would benefit from being able to survey coverage and quality of available genomic resources for selected groups of species from their field of interest. This would (i) aid project design, particularly in the context of comparative genomics analyses; (ii) simplify comparisons of the quality of their own data with that of existing assemblies; and (iii) provide a means to keep up to date with accumulating genomics resources relevant to their ongoing research projects.

To address these needs, we developed an automated analysis workflow that performs BUSCO assessments of assemblies for user-selected taxa from the NCBI, concurrently collating assembly metadata to build a catalogue of metrics in a taxonomically aware framework. To demonstrate the utility of standardized evaluations for a clade, we applied our workflow to the phylum Arthropoda, for which genome data are supporting research on a wide range of topics including their roles as pests and disease vectors [16]. Since sequencing of the fruit fly genome [17], sampling of arthropods has included ants and other Hymenoptera [18, 19], arachnids [20], beetles [21], butterflies and other Lepidoptera [22], flies and other Diptera [23, 24], hemipterans [25], and many others [26, 27]. Through efforts such as the i5k 5000 arthropod genomes initiative [28] and others, the arthropod genomics community has worked to overcome challenges in genome sequencing, assembly, and annotation [29–31]. Despite encompassing only a tiny fraction of all arthropod diversity and showing taxonomic biases in sampling, assemblies are accumulating rapidly and are now publicly available for hundreds of species [32, 33].

Our large-scale assessments allowed us to (i) survey the current taxonomic coverage and assembly quality across Arthropoda, (ii) examine how key assembly metrics relate to gene content completeness, (iii) quantify effects on assessment resolution using different BUSCO lineage datasets, (iv) compare the results of BUSCO v3 with the newer BUSCO v4, and (v) demonstrate how our workflow can be used to build a community resource. We provide the catalogue as an open resource for the arthropod genomics community, and the stand-alone open-source workflow for users to build their own catalogues tailored to the needs of their research communities. Enabling user-customizable, taxonomically aware, standardized, and updatable quality assessments of available genome assemblies will empower genomics data producers and users, as well as helping to prioritize species for genomic sequencing of Earth's biodiversity.

Results and Discussion

An automated workflow for assembly assessments

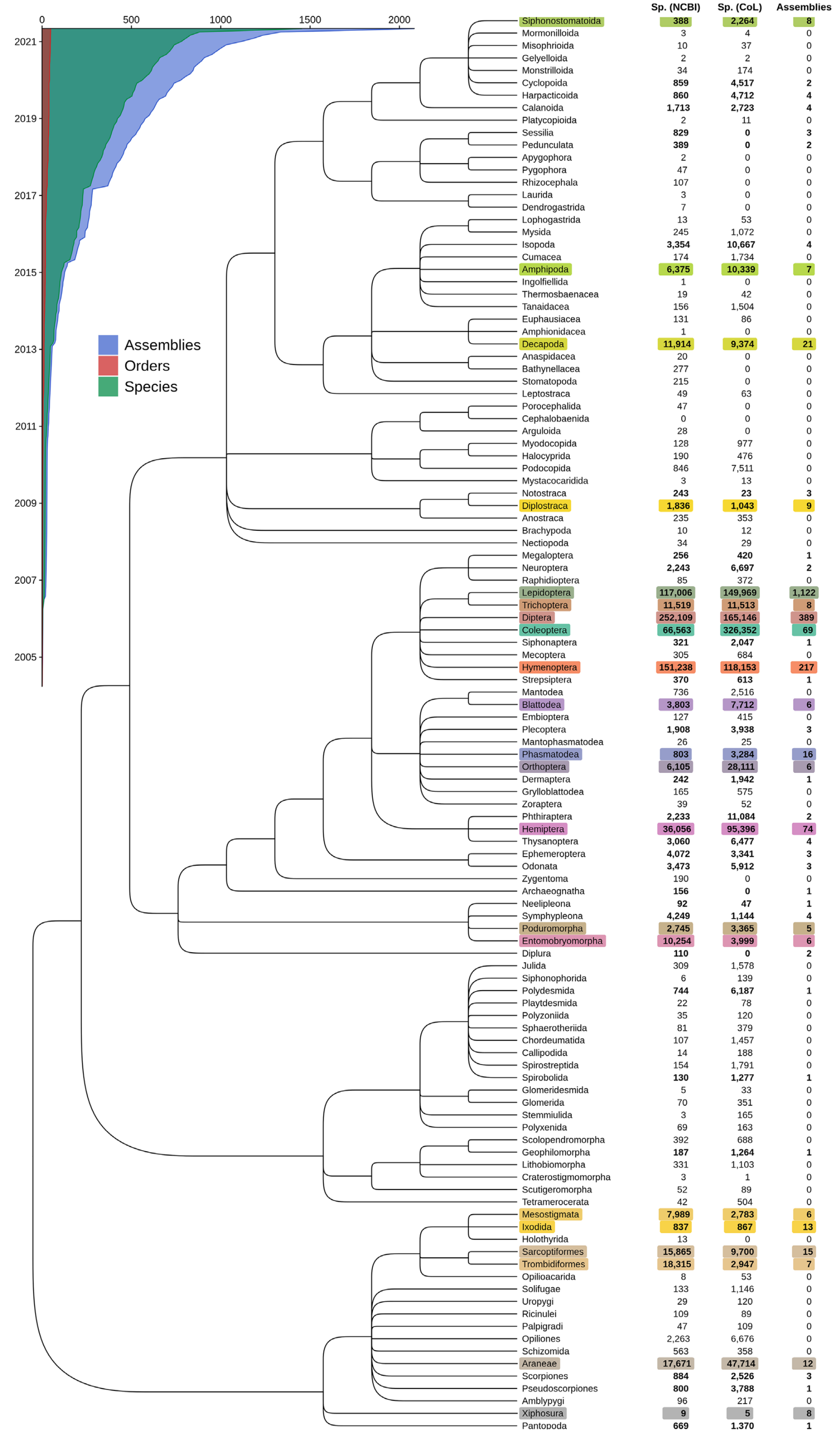
We developed an automated analysis workflow to build and maintain NCBI genome assembly assessment catalogues for selected taxa. This workflow performs the following steps: (i) query the NCBI GenBank Assembly database [11] to retrieve information about available assemblies and corresponding metadata for a user-defined taxonomic group; (ii) identify all relevant BUSCO lineages based on species taxonomy for each assembly; (iii) run BUSCO on each assembly using each relevant lineage dataset; (iv) generate a summary table that collates all BUSCO results with assembly metrics and metadata; and (v) generate an HTML/JavaScript interactive table containing all data from the summary (Supplementary Fig. S1). Assembly metadata are integrated into a summary file along with 5 metrics obtained from the results of running BUSCO on each assembly with each relevant lineage: the percentages of complete, complete single-copy, complete duplicated, fragmented, and missing BUSCOs. The workflow allows users to systematically assess all assemblies available at the NCBI for a given taxon of interest. Importantly, it is also designed to perform on-demand updates to assess assemblies added to NCBI GenBank since the last run. The final output provides all the information retrieved for each assembly in both JSON and tab-separated formats, and an HTML/JavaScript table is generated to display the data. This output is saved in a summary folder each time the workflow is run. The workflow is implemented using the Snakemake workflow management engine [34, 35], and all software dependencies are managed by the Conda package manager. It is fully automated and can be configured using a YAML file to specify the query to use for the NCBI Assembly database, BUSCO parameters, and the information to display in the output tables. The code and documentation are available from [36].

A survey of arthropod genome assembly resources

Applying the assembly assessment workflow to the phylum Arthropoda on 11 June 2021 resulted in the retrieval of a total of 2,083 assemblies from 1,387 species, providing a snapshot of the taxonomic coverage of available genome resources for arthropods at the NCBI. Of the ∼120 arthropod orders recognized by the NCBI Taxonomy database [37] or the Catalogue of Life [38], 48 are represented by ≥1 genome assembly, with 21 orders represented by ≥5 assemblies (Fig. 1). Currently available genome resources include 1,929 assemblies for 1,262 insect species and a further 154 assemblies for 125 other arthropod species. For Insecta, this is a doubling of the number of species since a November 2020 survey from Hotaling et al. [33]. Species with assemblies represent a ∼0.06% sampling from a total of ∼1 million described arthropod species (792,339 species records and 121 orders in the NCBI Taxonomy database on 10 August 2021; 1,126,288 extant species and 123 orders in the Catalogue of Life 2021–06-10 edition).

This survey highlights the sparsity and taxonomic imbalance of current species sampling, with 79.5% of species (83% of assemblies) belonging to only 3 orders: Lepidoptera—e.g., butterflies, moths (712 species, 1,122 assemblies), Diptera—e.g., flies (216 species, 389 assemblies), and Hymenoptera—e.g., ants, bees, wasps (175 species, 217 assemblies). Similar sampling biases were identified by the November 2020 survey of NCBI resources for Insecta [33], where order-level counts for 601 insect species from 20 orders were 28% Diptera, 20% Lepidoptera, and 27% Hymenoptera. Notably, while roughly one-third of insect orders are represented, only 5–10% of orders from other groups such as crustaceans, myriapods (e.g., centipedes, millipedes), and chelicerates (e.g., spiders, scorpions) have ≥1 assembly. Across Arthropoda, orders with the most sequenced species also show the highest proportions of sequenced versus Catalogue-of-Life–described species despite also being amongst the most species-rich clades: 0.063% sequenced species for Lepidoptera, 0.019% for Diptera, and 0.016% for Hymenoptera. An exception to this observation is Coleoptera—e.g., beetles, weevils, which has the highest number of described species to date with currently available genome assembly resources for only 0.007% of these species.

These uneven distributions likely reflect historical biases in research interests for dipterans, which include the model species *Drosophila melanogaster* and disease vectors like mosquitoes; for lepidopterans, which have been a model to study the genetic basis of complex traits and population genetics; and for hymenopterans, which include many well-studied social insects. While such



	Sp. (NCBI)	Sp. (CoL)	Assemblies
Siphonostomatoida	388	2,264	8
Mormonilloida	3	4	0
Misophrioida	10	37	0
Gelyelloida	2	2	0
Monstrilloida	34	174	0
Cyclopoida	859	4,517	2
Harpacticoida	860	4,712	4
Calanoida	1,713	2,723	4
Platycopioida	2	11	0
Sessilia	829	0	3
Pedunculata	389	0	2
Apygophora	2	0	0
Pygophora	47	0	0
Rhizocephala	107	0	0
Laurida	3	0	0
Dendrogastrida	7	0	0
Lophogastrida	13	53	0
Mysida	245	1,072	0
Isopoda	3,354	10,667	4
Cumacea	174	1,734	0
Amphipoda	6,375	10,339	7
Ingolfiellida	1	0	0
Thermosbaenacea	19	42	0
Tanaidacea	156	1,504	0
Euphausiacea	131	86	0
Amphionidacea	1	0	0
Decapoda	11,914	9,374	21
Anaspidacea	20	0	0
Bathynellacea	277	0	0
Stomatopoda	215	0	0
Leptostraca	49	63	0
Porocephalida	47	0	0
Cephalobaenida	0	0	0
Arguloida	28	0	0
Myodocopida	128	977	0
Halocyprida	190	476	0
Podocopida	846	7,511	0
Mystacocaridida	3	13	0
Notostraca	243	23	3
Diplostraca	1,836	1,043	9
Anostraca	235	353	0
Brachypoda	10	12	0
Nectiopoda	34	29	0
Megaloptera	256	420	1
Neuroptera	2,243	6,697	2
Raphidioptera	85	372	0
Lepidoptera	117,006	149,969	1,122
Trichoptera	11,519	11,513	8
Diptera	252,109	165,146	389
Coleoptera	66,563	326,352	69
Siphonaptera	321	2,047	1
Mecoptera	305	684	0
Hymenoptera	151,238	118,153	217
Strepsiptera	370	613	1
Mantodea	736	2,516	0
Blattodea	3,803	7,712	6
Embioptera	127	415	0
Plecoptera	1,908	3,938	3
Mantophasmatodea	26	25	0
Phasmatodea	803	3,284	16
Orthoptera	6,105	28,111	6
Dermaptera	242	1,942	1
Grylloblattodea	165	575	0
Zoraptera	39	52	0
Phthiraptera	2,233	11,084	2
Hemiptera	36,056	95,396	74
Thysanoptera	3,060	6,477	4
Ephemeroptera	4,072	3,341	3
Odonata	3,473	5,912	3
Zygentoma	190	0	0
Archaeognatha	156	0	1
Neelipleona	92	47	1
Symphypleona	4,249	1,144	4
Poduromorpha	2,745	3,365	5
Entomobryomorpha	10,254	3,999	6
Diplura	110	0	2
Julida	309	1,578	0
Siphonophorida	6	139	0
Polydesmida	744	6,187	1
Playtdesmida	22	78	0
Polyzoniida	35	120	0
Sphaerotheriida	81	379	0
Chordeumatida	107	1,457	0
Callipodida	14	188	0
Spirostreptida	154	1,791	0
Spirobolida	130	1,277	1
Glomeridesmida	5	33	0
Glomerida	70	351	0
Stemmiulida	3	165	0
Polyxenida	69	163	0
Scolopendromorpha	392	688	0
Geophilomorpha	187	1,264	1
Lithobiomorpha	331	1,103	0
Craterostigmomorpha	3	1	0
Scutigeromorpha	52	89	0
Tetramerocerata	42	504	0
Mesostigmata	7,989	2,783	6
Ixodida	837	867	13
Holothyrida	13	0	0
Sarcoptiformes	15,865	9,700	15
Trombidiformes	18,315	2,947	7
Opilioacarida	8	53	0
Solifugae	133	1,146	0
Uropygi	29	120	0
Ricinulei	109	89	0
Palpigradi	47	109	0
Opiliones	2,263	6,676	0
Schizomida	563	358	0
Araneae	17,671	47,714	12
Scorpiones	884	2,526	3
Pseudoscorpiones	800	3,788	1
Amblypygi	96	217	0
Xiphosura	9	5	8
Pantopoda	669	1,370	1

Figure 1: Available genome assembly resources across the arthropod phylogeny. The Arthropoda phylogeny from the US NCBI Taxonomy database shows the evolutionary relationships amongst 114 orders. Counts of described species (Sp.) within each order are shown from the NCBI (v.2021–06-11) and the Catalogue of Life (CoL, v.2021–06-10), alongside numbers of genome assemblies available from the NCBI Assembly database (accessed on 25 August 2021). Of the 114 orders recognized by both the NCBI and the CoL, 48 orders are represented by ≥1 genome assembly. The 21 orders with ≥5 assemblies are highlighted with distinct colours, which are maintained for cross-referencing in Figs 2–4. The inset shows the accumulation of assemblies, species, and orders submitted to the NCBI since 2005 (note that in the case of assembly updates only the latest submission dates are considered).

biases may persist owing to factors such as research priorities and ease of sampling, the balance should improve as the numbers and taxonomic spread of available arthropod genome assemblies continue to grow rapidly (Fig. 1, Inset). Surveying taxonomic representation in this way highlights the increasingly rapid accumulation of new genome assemblies at the NCBI, providing researchers with a comprehensive overview of the species coverage of available genomics resources for their taxa of interest.

Assessing the surveyed species data allows for phylum-wide comparisons of the contiguity and completeness of genome assemblies available at the NCBI. Focusing on the 21 orders with ≥5 assemblies, order representation is notably unbalanced and assembly quality metrics summarized with N50 lengths and BUSCO completeness scores vary greatly among and within orders (Fig. 2). Large differences between assembly and species counts are primarily driven in Lepidoptera by *Heliconius melpomene* (n = 42), *Junonia neildi* (n = 35), *Junonia evarete* (n = 32), and 6 other *Junonia* and *Heliconius* species with >10 assemblies, and in Diptera mainly by *D. melanogaster* (n = 26), *Drosophila simulans* (n = 12), and *Anopheles coluzzii* (n = 10). The 307 species with >1 assembly comprise distinct assembly submissions and not updates that result in new versions of existing submissions (in this case only the latest version is surveyed). Roughly half (142) of these species with multiple assemblies are represented by a chromosome-level assembly. Across all assemblies, those labelled as chromosome-level account for 12.3%, while a further 41.1% are labelled as scaffold-level assemblies, and the remaining 46.6% are contig-level (Supplementary Fig. S2).

Excluding Lepidoptera, which are skewed by a large number of poor-quality assemblies [39], median N50 lengths per order represented by ≥5 assemblies (shown in Fig. 2C) range from 11.6 kb for Sarcoptiformes (mites, 15 assemblies for 12 species) to 96.3 Mb for Xiphosura (horseshoe crabs, 8 assemblies for 4 species). The horseshoe crabs have large genomes of 1.7–2.2 Gb, for which concerted efforts have been successful in producing contiguous assemblies [40–43]. The mite genomes are all much smaller, with a median assembly span (total length) of just 88.5 Mb, where the latest assembly for the parasitic mite, *Sarcoptes scabiei*, provides an example of how long-read technologies are helping to improve available genomic resources [44].

Median BUSCO completeness scores per order represented by ≥5 assemblies for the Arthropoda lineage dataset (Fig. 2D) are less variable than the N50 lengths and, excluding Lepidoptera, range from 72.1% for Sarcoptiformes to >97% for Diplostraca (clam shrimps and waterfleas, 9 assemblies for 7 species), Blattodea (cockroaches and termites, 6 assemblies for 5 species), Diptera, and Hymenoptera. Although within-order distributions can be highly variable, all but 2 of the 21 orders (Sarcoptiformes and Trombidiformes mites) are represented by ≥1 assembly with >90% complete BUSCOs. These contiguity and completeness distributions include all available assemblies, i.e., not filtered by level (contig, scaffold, chromosome) or type (e.g., haploid, principal or alternate pseudohaplotype). The completeness of contig-level assemblies is expectedly lower than that of scaffold- or chromosome-level (Supplementary Fig. S2B) assemblies, and although alternate pseudohaplotype assemblies can achieve high BUSCO completeness scores, they are generally lower than for principal pseudohaplotypes (Supplementary Fig. S2C). Additional partitioning of the datasets by sequencing technologies, assembly algorithms, and so forth is feasible where the metadata labels are applied consistently, or after metadata curation as for previous assessments of insects that contrasted short- and long-read technologies [33]. These phylum-wide comparisons of the qualities of available genome assemblies highlight the unbalanced order-level species representation, as well as the variable levels of contiguity and completeness within and amongst arthropod orders.

Arthropod assembly contiguity, size, and completeness

With 2,083 assemblies exhibiting variable contiguities and sizes, the survey results provide the opportunity to examine expectations of how assembly contiguity and size relate to gene content completeness. Although long-read sequencing technologies are producing improved results [33], large genomes have often been challenging to assemble owing to expanded proportions of repetitive sequences [31]. Even for smaller genomes, repeats can hinder scaffolding of contigs, reducing contiguity and possibly adding undetermined gap regions to the assembly. Less contiguous assemblies are thus expected to have more genes split across scaffolds, or partially or completely missing, resulting in lower completeness scores [45].

The Earth BioGenome Project [2] criteria for a reference-quality assembly include obtaining a complete and single-copy BUSCO score >90% and having the majority of sequences assigned to chromosomes. While 828 of the assessed arthropod assemblies achieve a complete and single-copy BUSCO score >90%, only 229 of these are also labelled as chromosome-level assemblies. Indeed, comparing assembly N50 values with their completeness scores shows that obtaining >90% complete BUSCOs can be achieved across a wide range of contiguities (Fig. 3A). Recovery of >90% complete BUSCOs is observed for assemblies with N50s as low as 3.5 kb (*Tetragonula mellipes*, stingless bee, 92.1% complete) and 3.9 kb (*Chrysomya rufifacies*, blowfly, 97.4% complete). While some with N50s <10 kb are able to achieve >90% (n = 25) or 80–90% (n = 21) completeness, the vast majority of assemblies with such low contiguity levels achieve considerably lower BUSCO completeness scores than contiguous assemblies (i.e., N50 >10 kb). Among the latter, notable anomalies include 24 assemblies with N50s >10 kb that nonetheless all have completeness scores of <50%. One-third of these are labelled as alternate pseudohaplotypes, which offers an explanation for the low completeness levels because they likely represent collections of purged haplotigs. Others include improbably small assembly spans, e.g., *Sertania guttata* (butterfly, 30 Mb span of 628 Mb estimate) and *Dactylopius coccus* (scale insect, 18 Mb span of 386 Mb estimate), or high proportions of undetermined sequence, e.g., the brown recluse spider, *Loxosceles reclusa* (45% gaps). Biological complexity may also offer explanations, such as in the case of the Lord Howe Island stick insect, *Dryococelus australis* (N50 = 17.3 kb, 43.5% complete), a potentially hexaploid genome with an estimated size of 4.2 Gb that achieved an assembly span of 3.4 Gb [46].

The largest assemblies span >5 Gb, with the maximum reported for the Asian longhorned tick, *Haemaphysalis longicornis*, at 7.3 Gb, which shows 92% complete BUSCOs (Fig. 3B). The estimated genome size for this tick however is only 3.4 Gb, and a duplicated BUSCO score of 74.4% suggests that the applied assembly methods failed to collapse the alternative haplotypes. Indeed, an alternative assembly for this tick spans just 2.6 Gb and scores 89.5% complete and 2.1% duplicated BUSCOs. A handful of other large assemblies with high duplicated scores are annotated as being non-collapsed, but others with many duplicated BUSCOs are also likely diploid or partially diploid (Supplementary Fig. S3). The smallest reported genome size for an arthropod to date is that of the tomato russet mite, *Aculops lycopersici* (Trombidiformes), exceptionally streamlined at only 32.5 Mb [47].

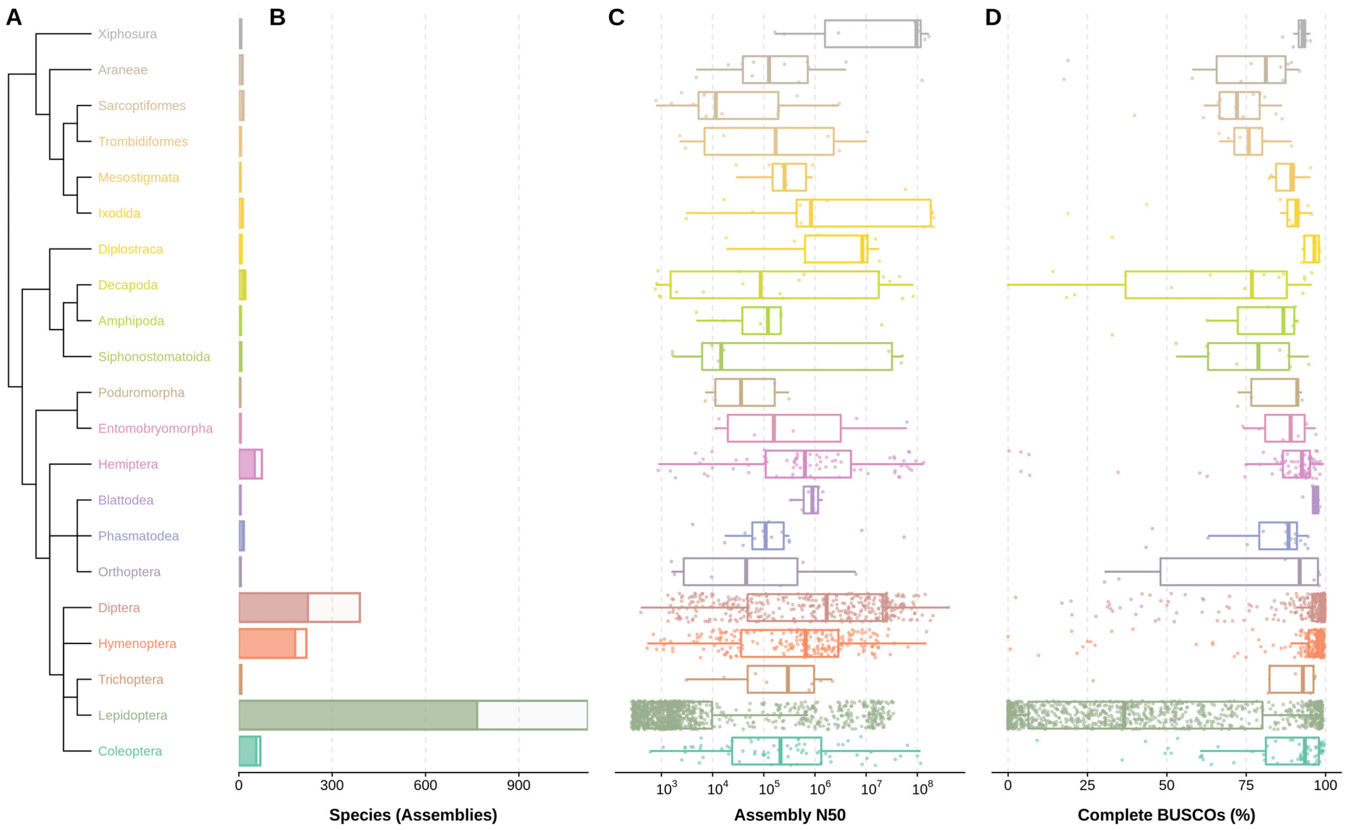


Figure 2: Order-level representation, contiguity, and completeness of 2,024 available assemblies for 1,326 arthropod species from the 21 orders with ≥5 assemblies. Data are presented only for orders with ≥5 assemblies available at the NCBI (2021–06-11). (A) Phylogenetic relationships of the 21 orders as resolved by the NCBI Taxonomy database. (B) Number of assemblies (entire bars) and unique species (dark fractions) retrieved from the NCBI Assembly database for each order. (C) Distribution of assembly NCBI scaffold N50 values (base pairs, log scale) for each order. (D) Distribution of BUSCO completeness (% of 1,013 BUSCOs) for the arthropod lineage dataset (arthropoda_odb10) for each order. Box plots show the median, first and third quartiles, and lower and upper extremes of the distribution (1.5 × IQR), and all values are overlaid as points to show the full distribution.

It achieves a Eukaryota completeness score of 83%, but only 67% Athropoda complete, which could reflect the evolutionary streamlining process but may also be related to challenges during gene prediction in such a gene-dense genome where genes have also experienced large-scale intron losses. The smallest assembly with a >80% Arthropoda completeness score is that of a grasshopper, *Xenocatantops brachycerus* (42 Mb, 92% complete); however, inspecting the metadata reveals this to be a transcriptome rather than a genome assembly [48]. Amongst the smallest true genome assemblies achieving >80% completeness are other Trombidiformes as well as Sarcoptiformes, e.g., the house dust mite *Dermatophagoides farinae* (54 Mb, 84% complete). Although there are fewer large assemblies spanning >1 Gb, across the full range of their sizes most achieve good completeness scores of >90%, indicating that sequencing technologies and assembly methods are able to overcome challenges often associated with large genomes.

Comparing assembly N50s and sizes with BUSCO duplicated scores (Supplementary Fig. S3) identifies several assemblies with high duplication levels. Some of these are labelled as "unresolved-diploid" assemblies, which explains these high duplication levels, but this mechanism to inform users about the non–strictly haploid status of certain assemblies is not widely nor consistently applied. Fragmented BUSCO scores (Supplementary Fig. S4) are expectedly higher for most of the less contiguous assemblies, highlighting those where many genes are likely split across 2 or more scaffolds. The survey results therefore provide the community with a comprehensive overview of genomic dataset qualities and of how contiguity and size relate to gene content completeness across currently available arthropod genome assemblies.

BUSCO dataset lineage and version comparisons

The reference BUSCO lineage datasets are defined at different taxonomic levels that capture sets of near-universal single-copy orthologues from OrthoDB [49] at ancient, intermediate, and younger nodes of the tree of life [8,9]. As duplication and loss events over evolutionary time erode the numbers of identifiable BUSCOs, datasets defined for more ancient lineages are smaller than for the younger ones, e.g., n = 255 for Eukaryota and n = 954 for Metazoa, versus n = 3,285 for Diptera and n = 13,780 for Primates (OrthoDB v10 datasets). An advantage of the smaller older lineage datasets is that compute runtimes are shorter because there are fewer individual genes to search for. The larger younger lineage datasets on the other hand offer greater resolution, meaning that scores are less affected by small differences in counts of complete, fragmented, or missing BUSCOs.

Our results provide the opportunity to compare the scores obtained using different lineage datasets for a large number of arthropod assemblies (Fig. 4). Comparing percentages of complete BUSCOs identified with the Eukaryota (n = 255) and the Arthropoda (n = 1,013) lineage datasets for a total of 1,977 arthropod assemblies shows highly linearly correlated scores, especially for the highest-scoring assemblies (Fig. 4A). For those scoring <80% there is a small but noticeable shift towards Arthropoda producing slightly higher scores than Eukaryota, indicating that proportion-

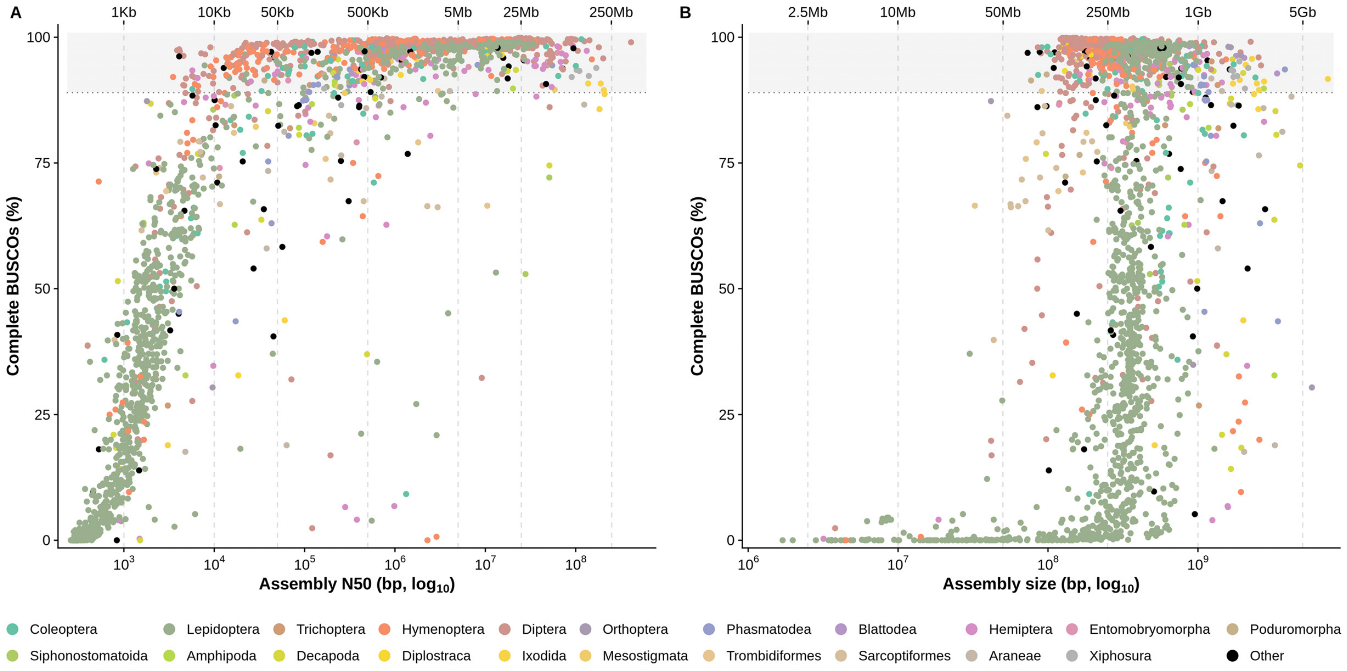


Figure 3: BUSCO completeness compared with assembly contiguity and size. Complete BUSCOs (in % of total BUSCOs for the arthropoda_odb10 dataset) are plotted against assembly N50 in bp (**A**) and assembly size in bp (**B**) for each assessed assembly. Both assembly N50 and assembly size are represented with a log scale. The colour of a point indicates the order of the sequenced species. Dotted lines indicate N50 values of 1, 10, 50, and 500 kb and 5 Mb in **A** and assembly size values of 50 Mb, 250 Mb, 1 Gb, and 5 Gb in **B**. BUSCO completeness scores >90% are highlighted with a grey background.

ately more of the larger set of Arthropoda BUSCOs can be recovered from lower-quality assemblies. Outlier points above the identity ($y = x$) axis suggest that the lower-resolution Eukaryota lineage dataset occasionally produces overestimates of completeness, where proportionately more of the smaller set of ancient Eukaryota BUSCOs are recovered. Similar trends are observed when comparing the Arthropoda results to the higher-resolution Insecta ($n = 1,367$) lineage dataset, with highly linearly correlated scores and occasional small overestimates of completeness using the Arthropoda lineage dataset (Supplementary Fig. S5A).

Comparing Arthropoda results to those from 4 insect order-level lineage datasets shows high agreements for the highest-scoring assemblies (Fig. 4B). For lower-scoring assemblies, results from applying the Lepidoptera and Hemiptera lineage datasets tend towards slightly higher scores than for Arthropoda. In contrast, using the Hymenoptera and Diptera lineage datasets generally produces lower completeness scores than for Arthropoda. These shifts could arise from the uneven representations of these orders in the 90-species Arthropoda lineage dataset, which is dominated by 20 hymenopterans and 15 dipterans, with only 9 species each for Lepidoptera and Hemiptera. The same trends are observed when comparing results from the order-level lineage datasets to those from the Insecta dataset (Supplementary Fig. S5B).

In addition to updates to the codebase, BUSCO v4 was released with updated lineage datasets based on orthology data from OrthoDB v10 [49], while BUSCO v3 used data from OrthoDB v9 [50]. Comparing completeness scores using the 2 Arthropoda datasets shows high levels of agreement for the highest-scoring assemblies with a consistent shift towards lower scores reported by BUSCO v4 for lower-quality assemblies (Fig. 4C). A similar pattern is observed when comparing results from the 2 Insecta datasets (Supplementary Fig. S5C). The Diptera comparisons on the other hand reveal some score variations, which nevertheless agree well over the full range of assembly qualities (Fig. 4D), similarly to results from the Hymenoptera datasets (Supplementary Fig. S5D). The different versions therefore produce generally consistent and comparable estimates of completeness, with a tendency for the OrthoDB-v10–based Arthropoda and Insecta datasets to report lower scores, especially for lower-quality assemblies. For objective quantitative comparisons it is thus necessary to assess assemblies using the same BUSCO versions, parameters, and lineage datasets, as presented here for the phylum-wide assessments of available arthropod genome assemblies.

The Arthropoda assembly assessment catalogue: A^{3}Cat

Running the workflow on the selected taxon of Arthropoda (NCBI:txid6656) produced the first version of the Arthropoda Assembly Assessment Catalogue (A^{3}Cat v.2021–06-11), demonstrating how the workflow can be used to build a community resource. The A^{3}Cat is provided as a searchable online table [51] (Arthropoda Assembly Assessment Catalog, RRID:SCR_021864) that makes it possible to browse and download the collated metadata and BUSCO assessment results for arthropod assemblies available from the NCBI ($n = 2,083$ for A^{3}Cat v.2021–06-11). Through simple text searches and/or applying query filters, users are able to quickly obtain downloadable overviews of the availability and quality of genome assembly resources for their arthropod taxa of interest. Without the computational burden of having to evaluate publicly available resources themselves, users can directly compare the assessments of their own assemblies with the precomputed results available from the A^{3}Cat. In addition, for version and parameter controlled like-for-like comparisons, a user-workflow is provided to compute quality metrics on user-provided

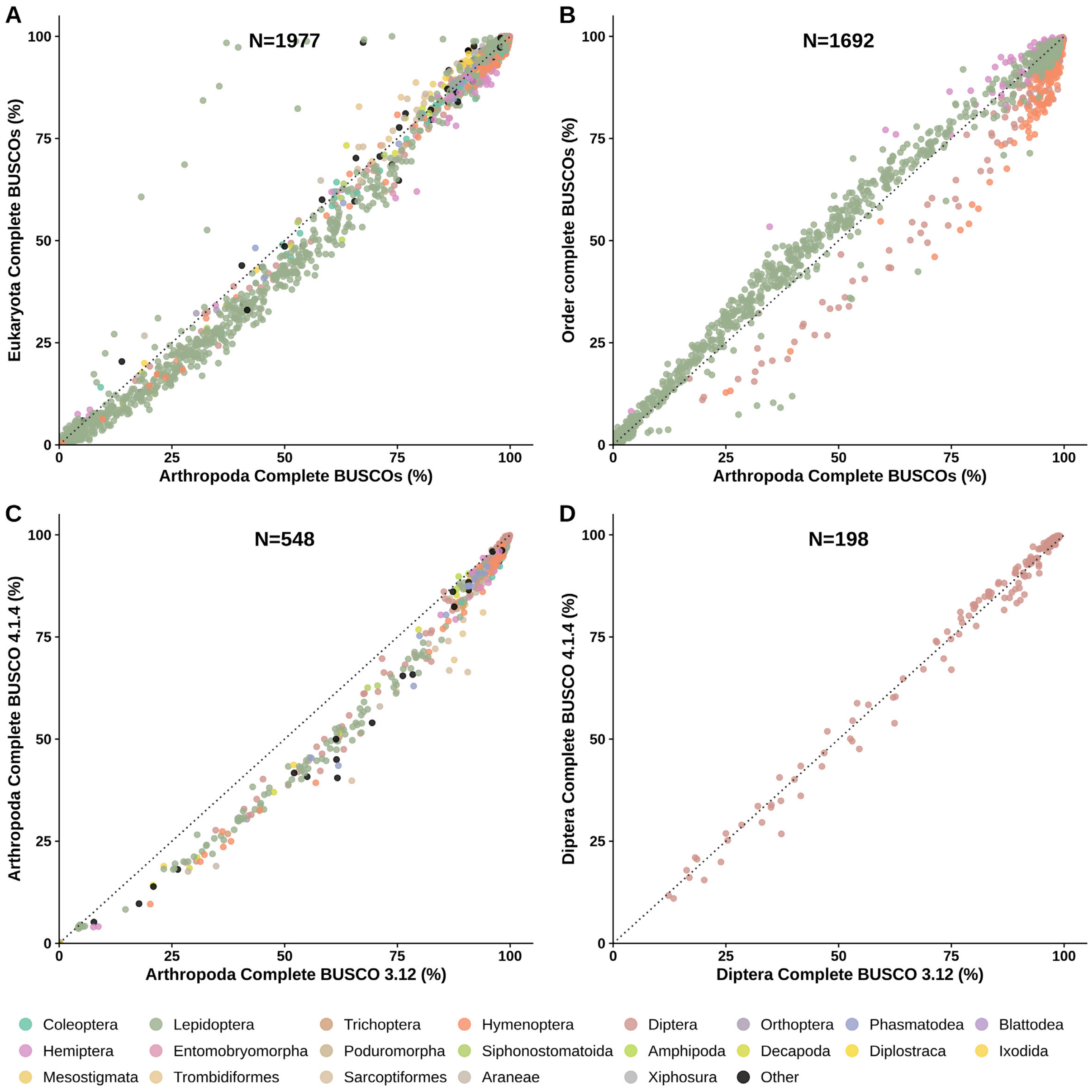


Figure 4: Comparisons of BUSCO lineage datasets and BUSCO versions. Congruence of BUSCO completeness scores is assessed by comparing results from (**A**) the Eukaryota (n = 255) and the Arthropoda (n = 1,013) lineage datasets, (**B**) the Arthropoda and 4 insect order-level lineage datasets (Hemiptera [n = 2,510], Hymenoptera [n = 5,991], Lepidoptera [n = 5,286], Diptera [n = 3,285]), and lineage datasets from BUSCO v4 (OrthoDBv10) and BUSCO v3 (OrthoDBv9) for (**C**) Arthropoda (odb9: n = 1,066) and (**D**) Diptera (odb9: n = 2,799). In each panel, the dotted lines show the identity (y = x).

assemblies and compare them with A^{3}Cat results for species from the same taxonomic clade (code and documentation are available from [52]).

Conclusions

Results from applying the assessment workflow to the phylum Arthropoda demonstrate the utility of building resources that provide a standardized overview of the current taxonomic coverage and quality of genome assembly resources available from the NCBI. The large-scale dataset also offers the opportunity to examine how widely used assembly metrics relate to BUSCO genespace completeness across a heterogeneous collection of genomes. Some anomalies point to errors or inconsistent use of metadata annotations where retractions or revisions would help to avoid misleading users about these resources. Furthermore, comparing results using different BUSCO datasets on large collections of assemblies reveals trends associated with using ancient (lower-resolution) or younger (higher-resolution) lineages, and datasets built for BUSCO v3 or v4. While congruence is high especially for high-scoring assemblies, truly objective comparisons require reporting of the BUSCO versions, parameters, and lineage datasets used. Our data will enable future large-scale comparisons with results from the recently released BUSCO v5, which includes a new genome assessment strategy that improves efficiency and runtimes [53]. Future workflow developments would aim to capture new metadata attributes made available from the NCBI such as summary information on repeat content, or computed locally, e.g., nucleotide compositions from *k*-mer analyses. The automated analysis workflow to build and maintain NCBI genome assembly assessment catalogues for selected taxa allows users to build updatable community resources, here exemplified with the A^{3}Cat, which facilitates surveying of species coverage and data quality for available arthropod assemblies and serves to guide ongoing and future genome generation initiatives.

Materials and Methods

Assembly selection and assessment workflow implementation

Accession numbers for all assemblies in the user-specified taxon are retrieved by querying the NCBI datasets API [54] with the ncbi-datasets-pylib library (version 12.3.0 in version 1.0 of a3cat-workflow) (Step 1 in Supplementary Fig. S1). For each assembly, the data package is downloaded to a temporary zip file using the "datasets" command-line utility (version 11.22.0 in version 1.0 of the a3cat-workflow). The nucleotide sequence and metadata are extracted from each data package with the ncbi-datasets-pylib library and stored as fasta and JSON files, respectively (Step 2 in Supplementary Fig. S1). For each assembly, complete taxonomic information is retrieved from the NCBI Taxonomy database [37] using the ete3 python module [55], version 3.1.2 in version 1.0 of the a3cat-workflow) and stored in a JSON file (Step 3 in Supplementary Fig. S1). Taxonomic information is used to determine all BUSCO lineage datasets relevant for each assembly (Step 4 in Supplementary Fig. S1). During this step, assemblies are filtered by size, scaffold N50, and a manual filter list to discard assemblies that are too short and/or fragmented to contain any BUSCOs; this is necessary because BUSCO returns an error if no BUSCOs are found. The completeness of each assembly is assessed using BUSCO in genome mode and all other settings to default (version 4.1.4 in version 1.0 of the a3cat-workflow) for each applicable lineage dataset (Step 5 in Supplementary Fig. S1). The results folder generated by BUSCO is saved as a compressed archive with the exception of the BLAST database (blast_db) and BLAST input sequences (<run_name>/blast_output/sequences). The full results table, missing BUSCO list, and short summary are also retained in the final output for convenience. Metadata retrieved from NCBI and BUSCO scores for all assemblies are aggregated into a JSON file that summarizes all the raw information retrieved and computed by the workflow (Step 6 in Supplementary Fig. S1). This JSON file is converted into a table with formatted headers stored in a tab-separated file where columns represent metadata and BUSCO scores and each line corresponds to an assembly (Step 7 in Supplementary Fig. S1). Finally, an interactive table is generated as an HTML page using the Data Tables JavaScript library [56] (version 1.10.24 in version 1.0 of the a3cat-workflow) (Step 8 in Supplementary Fig. S1). The entire workflow is implemented using the Snakemake workflow management engine [34, 35] and all software dependencies are managed by the Conda package manager; this implementation ensures that the workflow is portable and entirely reproducible. Parameters for each step of the workflow are specified in a YAML file and additional configuration files can be used to customize the table and HTML output. The code and documentation for the workflow are available from [36].

Assessment workflow deployment and data analyses

Results presented in this study were obtained by running version 1.0 of the a3cat-workflow on 11 June 2021. Species estimates were retrieved from the NCBI Taxonomy database using ete3 (version 3.1.2) on 21 August 2021 and from the Catalogue of Life version 2021–06-10. Phylogenetic trees were automatically generated from NCBI taxonomy data with ete3. BUSCO scores for version 4.1.4 were obtained directly from the output of a3cat-workflow, while scores for version 3.12 were obtained with a development release version of the workflow [57]. Figures were generated with ggplot2 version 3.3.5 [58] and ggtree version 3.0.1 [59] in R version 4.1.0 [60]. All data-related figures, numbers, and supplementary material were generated with a Snakemake workflow [35] available from [61] using Snakemake version 6.3.0.

Availability of Supporting Source Code and Requirements

Project name: The Arthropoda Assembly Assessment Catalogue Workflow
Project home page: https://gitlab.com/evogenlab/a3cat-workflow
Operating system: Platform independent
Programming language: Snakemake, Python
Other requirements: Snakemake, Conda
License: GPLv3
RRID:SCR_021864
biotools ID: arthropoda_assembly_assessment_catalogue

Data Availability

The data underlying this article are available in the NCBI Assembly Database at https://www.ncbi.nlm.nih.gov/assembly. An archival copy of the code and supporting data is also available via the *GigaScience* database GigaDB [62].

Additional Files

Supplementary Figure S1. Overview of the automated workflow for assembly assessments. The NCBI GenBank database is queried using the NCBI "datasets" python library **(1)** and assembly packages are downloaded with the "datasets" utility to obtain the genome sequence in a fasta file and metadata in a JSON file **(2)**. The complete taxonomy is retrieved from the NCBI taxonomy database for each assembly using ete3 **(3)** and used to determine relevant BUSCO lineage datasets **(4)**. BUSCO is then run with each lineage dataset on each assembly **(5)**, and BUSCO results are aggregated with taxonomy information and metadata into a single complete JSON summary file **(6)**. Finally, the summary is converted to a tab-separated table **(7)** and an HTML/Javascript searchable table is generated **(8)**.

Supplementary Figure S2. Accumulation over time and BUSCO completeness of contig-level, scaffold-level, or chromosome-level assemblies. (**A**) The cumulative numbers of assemblies labelled as contig-level, scaffold-level, and chromosome-level according to their submission dates at the NCBI Assembly database. (**B**) Distributions of BUSCO completeness scores for assemblies labelled as contig-level, scaffold-level, and chromosome-level at the NCBI Assembly database, and (**C**) those labelled as simply haploid, or distinguishing between the principal and alternate haplotypes. Box plots show the median, first and third quartiles, and lower and upper extremes of the distribution (1.5 × IQR), and all values are overlaid as points to show the full distribution.

Supplementary Figure S3. Proportion of duplicated BUSCOs compared with assembly contiguity and size. Duplicated BUSCOs (in % of total BUSCOs for the arthropoda_odb10 dataset) are plotted against assembly N50 in bp (**A**) and assembly size in bp (**B**) for each assessed assembly. Both assembly N50 and assembly size are represented with a log scale. The colour of a point indicates the order of the sequenced species. Dotted lines indicate N50 values of 1, 10, 50, and 500 kb and 5 Mb in panel **A** and assembly size values of 50 Mb, 250 Mb, 1 Gb, and 5 Gb in panel **B**.

Supplementary Figure S4. Proportion of fragmented BUSCOs compared with assembly contiguity and size. Fragmented BUSCOs (in % of total BUSCOs for the arthropoda_odb10 dataset) are plotted against assembly N50 in bp (**A**) and assembly size in bp (**B**) for each assessed assembly. Both assembly N50 and assembly size are represented with a log scale. The colour of a point indicates the order of the sequenced species. Dotted lines indicate N50 values of 1, 10, 50, and 500 kb and 5 Mb in panel **A** and assembly size values of 50 Mb, 250 Mb, 1 Gb, and 5 Gb in panel **B**.

Supplementary Figure S5. BUSCO dataset comparisons for Insecta and Hymenoptera. Congruence of BUSCO completeness scores is assessed by comparing results from the Arthropoda (n = 1,013) and Insecta (n = 1,367) lineage datasets (**A**), the Insecta and 4 insect order-level lineage datasets (Hemiptera [n = 2,510], Hymenoptera [n = 5,991], Lepidoptera [n = 5,286], Diptera [n = 3,285]) (**B**), and lineage datasets from BUSCO v4 (OrthoDBv10) and BUSCO v3 (OrthoDBv9) for Insecta (odb9: n = 1,658) (**C**) and Hymenoptera (odb9: n = 4,415) (**D**). Dotted lines represent the identity (y = x).

Abbreviations

API: application programming interface; BLAST: Basic Local Alignment Search Tool; bp: base pairs; BUSCO: Benchmarking Universal Single-Copy Orthologs; Gb: gigabase pairs; IQR: interquartile range; JSON: JavaScript Object Notation; kb: kilobase pairs; Mb: megabase pairs; NCBI: National Center for Biotechnology Information.

Competing Interests

The authors declare that they have no competing interests.

Funding

This research was supported by Novartis Foundation for medical-biological research grant No. 18B116 and Swiss National Science Foundation grants PP00P3_170664 and PP00P3_202669 to R.M.W.

Authors' Contributions

R.M.W. conceived the study. R.F. developed the workflows and performed the analyses. Both authors wrote the manuscript and read and approved the manuscript.

Acknowledgements

The authors thank Sagane Dind, Giulia Campli, Livio Ruzzante, and Antonin Thiébaut, as well as the reviewers, for providing useful suggestions for improvements and valuable feedback on the workflow and the manuscript.

References

1. Richards, S. It's more than stamp collecting: how genome sequencing can unify biological research. *Trends Genet* 2015;**31**(7):411–21.
2. Lewin, HA, Robinson, GE, Kress, WJ, *et al.* Earth BioGenome Project: sequencing life for the future of life. *Proc Natl Acad Sci U S A* 2018;**115**(17):4325–33.
3. Zoonomia Consortium. A comparative genomics multitool for scientific discovery and conservation. *Nature* 2020;**587**(7833):240–5.
4. Feng, S, Stiller, J, Deng, Y, *et al.* Dense sampling of bird diversity increases power of comparative genomics. *Nature* 2020;**587**(7833):252–7.
5. Thrash, A, Hoffmann, F, Perkins, A. Toward a more holistic method of genome assembly assessment. *BMC Bioinformatics* 2020;**21**(S4):249.
6. Dohmen, E, Kremer, LPM, Bornberg-Bauer, E, *et al.* DOGMA: domain-based transcriptome and proteome quality assessment. *Bioinformatics* 2016;**32**(17):2577–81.
7. Kemena, C, Dohmen, E, Bornberg-Bauer, E. DOGMA: a web server for proteome and transcriptome quality assessment. *Nucleic Acids Res* 2019;**47**(W1):W507–10.
8. Simão, FA, Waterhouse, RM, Ioannidis, P, *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;**31**(19):3210–2.
9. Waterhouse, RM, Seppey, M, Simão, FA, *et al.* BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* 2018;**35**(3):543–8.
10. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 2019;**47**(D1):D506–15.
11. Sayers, EW, Beck, J, Bolton, EE, *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2021;**49**(D1):D10–7.
12. Ewels, P, Magnusson, M, Lundin, S, *et al.* MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**(19):3047–8.
13. Challis, R, Richards, E, Rajan, J, *et al.* BlobToolKit – Interactive Quality Assessment of Genome Assemblies. *G3 (Bethesda)* 2020;**10**(4):1361–74.
14. Waterhouse, RM, Tegenfeldt, F, Li, J, *et al.* OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res* 2013;**41**(D1):D358–65.
15. Zdobnov, EM, Kuznetsov, D, Tegenfeldt, F, *et al.* OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 2021;**49**(D1):D389–93.
16. Childers, AK, Geib, SM, Sim, SB, *et al.* The USDA-ARS Ag100Pest Initiative: high-quality genome assemblies for agricultural pest arthropod research. *Insects* 2021;**12**(7):626.
17. Adams, MD. The genome sequence of *Drosophila melanogaster*. *Science* 2000;**287**(5461):2185–95.
18. Favreau, E, Martínez-Ruiz, C, Rodrigues Santiago, L, *et al.* Genes and genomic processes underpinning the social lives of ants. *Curr Opin Insect Sci* 2018;**25**:83–90.
19. Branstetter, MG, Childers, AK, Cox-Foster, D, *et al.* Genomes of the Hymenoptera. *Curr Opin Insect Sci* 2018;**25**:65–75.
20. Garb, JE, Sharma, PP, Ayoub, NA. Recent progress and prospects for advancing arachnid genomics. *Curr Opin Insect Sci* 2018;**25**:51–57.
21. McKenna, DD. Beetle genomes in the 21st century: prospects, progress and priorities. *Curr Opin Insect Sci* 2018;**25**: 76–82.

22. Triant, DA, Cinel, SD, Kawahara, AY. Lepidoptera genomes: current knowledge, gaps and future directions. *Curr Opin Insect Sci* 2018;**25**:99–105.
23. Wiegmann, BM, Richards, S. Genomes of Diptera. *Curr Opin Insect Sci* 2018;**25**:116–24.
24. Ruzzante, L, Reijnders, M, Waterhouse, RM. Of genes and genomes: mosquito evolution and diversity. *Trends Parasitol* 2019;**35**(1):32–51.
25. Panfilio, KA, Angelini, DR. By land, air, and sea: hemipteran diversity through the genomic lens. *Curr Opin Insect Sci* 2018;**25**:106–15.
26. González, VL, Devine, AM, Trizna, M, *et al.* Open access genomic resources for terrestrial arthropods. *Curr Opin Insect Sci* 2018;**25**:91–98.
27. Richards, S, Childers, A, Childers, C. Editorial overview: Insect genomics: Arthropod genomic resources for the 21st century: It only counts if it's in the database! *Curr Opin Insect Sci* 2018;**25**:iv–vii.
28. i5K Consortium. The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered* 2013;**104**(5):595–600.
29. Brown, SJ, Tagu, D. Editorial overview: Insect genomics: How to sequence five thousand insect genomes? *Curr Opin Insect Sci* 2015;**7**:iv–v.
30. Waterhouse, RM. A maturing understanding of the composition of the insect gene repertoire. *Curr Opin Insect Sci* 2015;**7**:15–23.
31. Li, F, Zhao, X, Li, M, *et al.* Insect genomes: progress and challenges. *Insect Mol Biol* 2019;**28**(6):739–58.
32. Hotaling, S, Kelley, JL, Frandsen, PB. Aquatic insects are dramatically underrepresented in genomic research. *Insects* 2020;**11**(9):601.
33. Hotaling, S, Sproul, JS, Heckenhauer, J, *et al.* Long-reads are revolutionizing 20 years of insect genome sequencing. *Genome Biol Evol* 2021;**13**(8):doi:10.1093/gbe/evab138.
34. Köster, J, Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012;**28**(19):2520–2.
35. Mölder, F, Jablonski, KP, Letcher, B, *et al.* Sustainable data analysis with Snakemake. *F1000Res* 2021;**10**:33.
36. Feron, R. a3cat-workflow. https://gitlab.com/evogenlab/a3cat-workflow. Accessed 21 October 2021.
37. Schoch, CL, Ciufo, S, Domrachev, M, *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020;**2020**:doi:10.1093/database/baaa062.
38. Roskov, Y, Ower, G, Orrell, T, *et al.* Catalogue of Life - 2019 Annual Checklist. 2020. http://www.catalogueoflife.org/annual-checklist/2019/info/ac. Accessed 13 May 2020.
39. Ellis, EA, Storer, CG, Kawahara, AY. De novo genome assemblies of butterflies. *Gigascience* 2021;**10**(6):doi:10.1093/gigascience/giab041.
40. Zhou, Y, Liang, Y, Yan, Q, *et al.* The draft genome of horseshoe crab *Tachypleus tridentatus* reveals its evolutionary scenario and well-developed innate immunity. *BMC Genomics* 2020;**21**(1):137.
41. Shingate, P, Ravi, V, Prasad, A, *et al.* Chromosome-level assembly of the horseshoe crab genome provides insights into its genome evolution. *Nat Commun* 2020;**11**(1):2322.
42. Shingate, P, Ravi, V, Prasad, A, *et al.* Chromosome-level genome assembly of the coastal horseshoe crab (*Tachypleus gigas*). *Mol Ecol Resour* 2020;**20**(6):1748–60.
43. Nong, W, Qu, Z, Li, Y, *et al.* Horseshoe crab genomes reveal the evolution of genes and microRNAs after three rounds of whole genome duplication. *Commun Biol* 2021;**4**(1):83.
44. Korhonen, PK, Gasser, RB, Ma, G, *et al.* High-quality nuclear genome for *Sarcoptes scabiei*—A critical resource for a neglected parasite. *PLoS Negl Trop Dis* 2020;**14**(10):e0008720.
45. Waterhouse, RM, Seppey, M, Simão, FA, *et al.* Using BUSCO to assess insect genomic resources. *Methods Mol Biol* 2019;**1858**:59–74.
46. Mikheyev, AS, Zwick, A, Magrath, MJL, *et al.* Museum genomics confirms that the Lord Howe Island stick insect survived extinction. *Curr Biol* 2017;**27**(20):3157–3161.e4.
47. Greenhalgh, R, Dermauw, W, Glas, JJ, *et al.* Genome streamlining in a minute herbivore that manipulates its host plant. *eLife* 2020;**9**:doi:10.7554/eLife.56689.
48. Zhao, L, Zhang, X, Qiu, Z, *et al.* De novo assembly and characterization of the *Xenocatantops brachycerus* transcriptome. *Int J Mol Sci* 2018;**19**(2):520.
49. Kriventseva, EV, Kuznetsov, D, Tegenfeldt, F, *et al.* OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 2019;**47**(D1):D807–11.
50. Zdobnov, EM, Tegenfeldt, F, Kuznetsov, D, *et al.* OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* 2017;**45**(D1):D744–9.
51. Waterhouse, RM. a3cat. https://rmwaterhouse.org/a3cat. Accessed 21 October 2021.
52. Feron, R. a3cat-user-workflow. GitLab. https://gitlab.com/evogenlab/a3cat-user-workflow. Accessed 21 October 2021.
53. Manni, M, Berkeley, MR, Seppey, M, *et al.* BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol Biol Evol* 2021;**38**(10):4647–54.
54. NCBI Datasets. https://www.ncbi.nlm.nih.gov/datasets. Accessed 21 October 2021.
55. Huerta-Cepas, J, Serra, F, Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol* 2016;**33**(6):1635–8.
56. DataTables | Table plug-in for jQuery. https://datatables.net. Accessed 21 October 2021.
57. Feron, R. paper-busco-v3 · Waterhouse Lab /a3cat-workflow. GitLab. https://gitlab.com/evogenlab/a3cat-workflow/-/releases/paper-busco-v3. Accessed 21 October 2021.
58. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer; 2009.
59. Yu, G, Smith, DK, Zhu, H, *et al.* ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017;**8**(1):28–36.
60. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
61. Feron, R. paper-a3cat. GitLab. https://gitlab.com/evogenlab/paper-a3cat. Accessed 21 October 2021.
62. Feron, R, Waterhouse, R. Supporting data for "Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes." *GigaScience Database* 2022. http://dx.doi.org/10.5524/100974.

Appendix 2: Review publication in Current Opinion in Insect Science

Publication

Exploring new genomic territories with emerging model insects

Feron and Waterhouse. Curr Opin Insect Sci. 2022 Jun;51:100902. doi:10.1016/j.cois.2022.100902. Epub 2022 Mar 14. PMID: 35301165

Author contribution

I performed the analyses, produced the figures, and contributed to writing and editing the manuscript including reviewing the literature to identify examples of new emerging model systems in insect genomics benefiting from new high-quality genomics resources.

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Insect Science

Exploring new genomic territories with emerging model insects

Romain Feron and Robert M Waterhouse

Improvements in reference genome generation for insects and across the tree of life are extending the concept and utility of model organisms beyond traditional laboratory-tractable supermodels. Species or groups of species with comprehensive genome resources can be developed into model systems for studying a large variety of biological phenomena. Advances in sequencing and assembly technologies are supporting these emerging genome-enabled model systems by producing resources that are increasingly accurate and complete. Nevertheless, quality controls including assessing gene content completeness are required to ensure that these data can be included in expanding catalogues of high-quality references that will greatly advance understanding of insect biology and evolution.

Address
Department of Ecology and Evolution, University of Lausanne, and the Swiss Institute of Bioinformatics,1015 Lausanne, Switzerland

Corresponding author:
Robert M Waterhouse (robert.waterhouse@unil.ch)

Current Opinion in Insect Science 2022, **51**:100902

This review comes from a themed issue on **Insect genomics**

Edited by **Shuji Shigenobu**

For complete overview about the section, refer “Insect genomics (February 2022)”

Available online 14th March 2022

https://10.1016/j.cois.2022.100902

Introduction

Model organisms can be described as non-human species that are studied to advance the understanding of biological phenomena, with traditional model species being easily bred in the laboratory and amenable to experimental manipulation [1]. The common ancestry of living organisms means that insights from such models also inform knowledge of molecular and genetic mechanisms underlying common biological functions across the tree of life. Representing insects is the renowned model, the fruit fly *Drosophila melanogaster*, with groundbreaking work on fields from genetics and heredity to behaviour, physiology, development, immunity, and countless others [2]. A major contributing factor to the success of *Drosophila* as a versatile model over the last two decades was the establishment of a reference genome assembly and its functional genomic element annotations [3]. Developing new models with reference genomes and experimental tools analogous to those available for *Drosophila* can be challenging, but is important for diversifying the systems we use to learn about organismal biology [4,5]. Currently, substantial advances in sequencing technologies mean that it can be more readily feasible to generate a high-quality genome for a new species than it is to rear in the laboratory. This genomics revolution is opening up a whole new set of possibilities considering a shift from the traditional model organism to the concept of species or groups of species that offer the ability to develop new model systems for studying a large variety of biological phenomena at many different levels [6,7].

Conserved orthologues help gauge gene content completeness of accumulating genome resources

Recent surveys of the current status of available genome resources for insects focus on taxonomic representation, assembly quality metrics, gene content completeness, and sequencing technology use [8–10]. These highlight the continued rapid accumulation since previous surveys, for example [11,12], and show current biases in species sampling with several insect orders still lacking publicly available resources. Notably, long-read data, for example, from approaches developed by Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT) are helping to improve assembly contiguity and produce more complete and accurate representations of new and upgraded insect genomes. For these resources to support the development of emerging model systems, they need to be of the highest possible quality, not only in terms of assembly statistics but also with respect to gene content representation.

The need to assess quality in terms of expected gene content prompted the proposal of Benchmarking Universal Single-Copy Orthologues (BUSCOs) [13]. BUSCO relies on the expectation that single-copy orthologues present in most species within a taxonomic lineage should be identifiable in any new genome from a species in the same clade. The BUSCO lineage datasets are built by identifying near-universal single-copy orthologues from the OrthoDB orthology resource [13–15]. Using these to evaluate assemblies starts with BUSCO sequence searches to guide gene predictions, then orthology classifications identify complete, duplicated, or fragmented BUSCOs. The numbers of identifiable BUSCOs provide an indication of gene content completeness based on expected subsets of evolutionarily conserved genes for a given lineage. High completeness scores thereby imply that a genome assembly confidently represents the complete gene repertoire.

Development of BUSCO assessments resulted in an initial implementation [16] with three lineage datasets relevant for insects (Eukaryota, Metazoa, and Arthropoda) based on the orthology data from OrthoDB v8 [14]. Subsequent updates in BUSCO v3 [17] provided four more lineage datasets within the Arthropoda (Insecta n=1658 BUSCOs, Endopterygota n=2442, Diptera n=2799, Hymenoptera n=4415), using orthologues from OrthoDB v9 [15]. The latest BUSCO releases [18] now provide additional insect-relevant lineage datasets for Hemiptera (n=2510) and Lepidoptera (n=5286), from OrthoDB v10 [19]. The assessments provide measures of data quality, and protocols for applications to insect genomic data [20] and for wider uses [21] help users to identify the best available genomic resources. BUSCO completeness is also recognised as an important quality check of resources for new model systems and for cataloguing eukaryotic genomic biodiversity, for example, the Earth BioGenome Project (EBP) recommendations on standards for genome generation include achieving recovery of more than 90% single-copy conserved genes [22].

Using results from the Arthropoda Assembly Assessment Catalogue (A^{3}Cat) [10,23] to survey BUSCO completeness of insect genome assemblies deposited at the United States National Center for Biotechnology Information (NCBI) shows that while many do meet EBP's recommendations, quality in terms of gene content completeness still varies dramatically (Figure 1). Thus, while the NCBI may currently offer more than 2500 assemblies for insects, fewer than half of these achieve a complete and single-copy BUSCO score > 90% and most do not yet reach the EBP's standard of having the majority of sequences assigned to chromosomes. Notably, however, accuracy-enhanced long-read technologies together with scaffolding approaches such as high-throughput chromatin conformation capture (Hi-C) are more consistently producing high-quality new genome resources, which are greatly expanding the possibilities for developing new insect model systems.

Emerging insect model systems are supported by high-quality genome resources

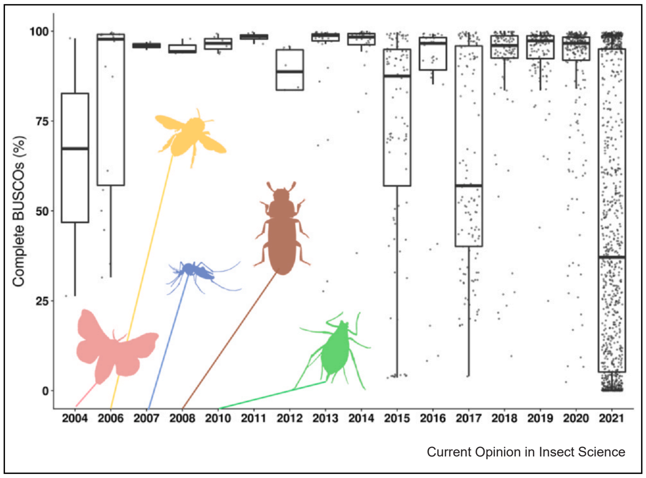
Advances in taxonomic sampling of insects for genome sequencing have been reviewed for ants and other Hymenoptera [24,25], hemipterans [26], beetles [27], flies and other Diptera [28,29], butterflies and other Lepidoptera [30], and many others [9,11,31]. Here, we focus on a selection of recent examples of high-quality genomics resources (Table 1) that are supporting the use of new species or groups of species to develop and expand emerging model systems that help advance understanding of insect biology and evolution.

Mayflies have long been the focus of many ecological studies, and together with dragonflies and damselflies they form the sister group to all other winged insect lineages. Recent establishment of a continuous culture system of the *Cloeon dipterum* mayfly [32] allows for comprehensive life-stage and tissue sampling for detailed transcriptional profiling. Combining short reads with ONT sequencing data enabled the assembly of its relatively compact genome of 180 Megabasepairs (Mbp) in 1395 scaffolds with 96%–97% complete BUSCOs (Table 1), and annotated with 16357 protein-coding genes. These resources lay the foundations for investigating genomic adaptations to aquatic and aerial life and the origin of insect wings in this emerging model system [32].

Combining long reads with Hi-C data is proving to be an effective approach for generating chromosome-level assemblies. This has been recently demonstrated by Sun et al. [33] for 5 of 17 new high-quality bumblebee genomes (Table 1), where comparisons revealed how the 25-chromosome karyotype of the social parasite species derived from the ancestral karyotype of 18 chromosomes. These resources are helping to set up the *Bombus* genus as a new model for quantifying genetic and genomic variations underlying important ecological and behavioural traits of key pollinators. Along with other new bumblebee reference genomes [34,35], they also offer opportunities to explore the genetic factors influencing the plastic and adaptive responses impacting insect resilience to climate change [36].

Rearrangements like those observed for the social parasite bumblebees seem to be infrequent in some well-studied groups such as Diptera and Lepidoptera, where global genome architectures are generally conserved. Therefore, models from other diverse insect groups are needed to investigate the different modes of genome structure evolution. Indeed, analyses of high-quality chromosome-level assemblies of aphids (Table 1) show that their autosomes have undergone dramatic reorganisations in contrast to their sex chromosomes, where gene content of the X chromosome has remained highly stable [37,38]. As a model system to investigate the evolution of resistance to insecticides, reference-quality aphid genomes are also enabling comprehensive assessments of within-species variation to understand genomic responses to strong selective forces [39].

Figure 1



BUSCO completeness of insect genome assemblies deposited at the United States NCBI. The boxplots show distributions per year of the percentage of complete BUSCOs assessed using the Arthropoda lineage dataset for insect assemblies available from the NCBI Assembly database. The first decade is characterised by a slowly increasing number of genome assembly releases, usually for what are regarded as some of the most charismatic and well-studied model insect species, and mostly showing high BUSCO completeness. The subsequent years are characterised by a much faster rate of growth in the numbers of genome assembly releases, accompanied by large variations in quality in terms of gene content completeness. The large numbers of low-completeness assemblies deposited in 2017 and 2021 comprise mainly those for the lepidopteran species. Insect silhouettes depict, from left to right: *Bombyx mori* silkmoth, *Apis mellifera* honey bee, *Aedes aegypti* mosquito, *Tribolium castaneum* beetle, and *Acyrthosiphon pisum* pea aphid, linked to the year their genome was first published. Boxplots show the median, first and third quartiles, and lower and upper extremes of the distribution (1.5 x interquartile range). Data are sourced from the Arthropoda Assembly Assessment Catalogue (A^{3}Cat) [10, 23]; data for 2021 are shown only for assemblies available up to June 11.

The pea aphid was one of the first insects to be sequenced and has served as a valuable model for understanding genomic consequences of host-symbiont interactions. However, genomic resources for new systems are needed to explore the many types of endosymbioses found across different insects. The genome of the rice weevil, *Sitophilus oryzae*, is not yet assembled to the chromosome level but shows high BUSCO completeness (Table 1), thereby providing a confident basis from which to investigate how key metabolic processes might be partitioned between the host and the endosymbiont [40]. Quality and completeness are also particularly critical when tracing cases of horizontal gene transfer, for example, duplicated bacterial-origin mannosidases in the 1150 Mbp genome assembly of the stink bug *Halyomorpha halys* [41], and bacterial cell wall hydrolase genes acquired by Coccinellinae ladybird beetles identified in the high-quality genome of *Cryptolaemus montrouzieri* [42].

Among the most well-known of the Coccinellinae, the harlequin ladybird *Harmonia axyridis* is widely considered to be one of the world's most invasive insects. Many insects are, or have the potential to become, invasives that can cause great damage to natural ecosystems or agricultural crops. Accumulating genomic resources from a variety of insect groups are helping to diversify the models used to study invasion biology and potentially develop new genetic control measures. Hi-C data helped to build a chromosome-level assembly for the two-spot harlequin morph, but with lower BUSCO completeness than the earlier Hi-C scaffolding [43] (Table 1). These data, along with assemblies for other morphs, for example [44], also offer new opportunities to develop the use of these ladybirds, which display more than 200 described colour forms, as an important model system for investigating the genetics of colour pattern polymorphisms [45,46].

Being laboratory tractable is a key feature of the most versatile model species. For example, the painted lady butterfly, *Vanessa cardui*, can be easily reared in the laboratory and is amenable to CRISPR/Cas9 genome editing, making this widespread, generalist species with complex wing patterns an excellent model. The genome assembly, recently upgraded to chromosome level [47], with transcriptomics data from multiple tissues and developmental stages provides the framework to employ genetic manipulations and functional genomics data for studying migration, host-plant coevolution, and colour patterning [48]. CRISPR/Cas9 has also been established for the tea geometrid moth, *Ectropis grisescens*, which, along with its relevance as an agricultural pest, presents an interesting system for studying insect interactions with plant allelochemicals as well as shape and colour adaptations for effective camouflage. Hi-C scaffolding of PacBio data placed 97.8% of the assembly on 31 chromosomes with an assembly span of 785 Mbp (Table 1) and 18746 annotated protein-coding genes. The genome maintains the ancestral lepidopteran karyotype (n=31), and separate resequencing of male (ZZ) and female (ZW) individuals allowed for the identification of the Z chromosome and several W candidate scaffolds [49].

While still often challenging, long reads are proving particularly useful for assembling such repeat-rich insect sex chromosomes. For example, the *Pieris macdunnoughii* assembly (Table 1) was built using ONT long reads, where polishing with additional short-read data increased complete lepidopteran BUSCOs by almost 3%. Comparing the resolved sex chromosomes in *Pieris* butterflies of European and North American lineages shows that the fusion event that created the neo-Z chromosome occurred before their divergence [50]. These genome resources support this emerging model system for studying maladaptation in plant–insect interactions, where the North American butterflies lay

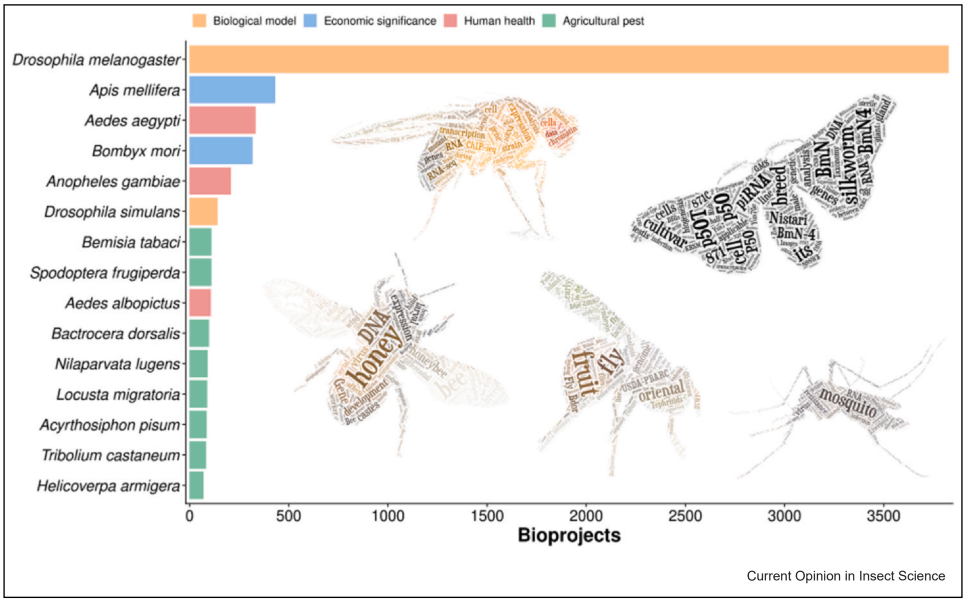
Table 1

Selected examples of emerging models supported by high-quality genome resources.

Taxon	Assembly Size (Mbp)	Scaffold N50 (Mbp)	Insecta BUSCO % C,[S,D],F,M	Arthropoda BUSCO % C,[S,D],F,M
Mayfly: *Cloeon dipterum*	180	0.46	96.1,[94.1,2.0],0.9,3.0	97.2,[95.0,2.2],1.2,1.6
Bumblebees: *Bombus haemorrhoidalis*	241	15.09	99.7,[99.4,0.3],0.2,0.1	99.4,[99.3,0.1],0.3,0.3
Bumblebees: *Bombus ignitus*	243	15.19	98.3,[98.1,0.2],0.7.1.0	97.6,[97.5,0.1],1.3,1.1
Bumblebees: *Bombus turneri*	243	9.70	99.6,[99.3,0.3],0.2,0.2	99.2,[99.2,0.0],0.5,0.3
Bumblebees: *Bombus breviceps*	248	14.71	99.6,[99.4,0.2],0.1,0.3	99.1,[99.1,0.0],0.4,0.5
Bumblebees: *Bombus pyrosoma*	255	15.22	99.7,[99.5,0.2],0.1,0.2	99.6,[99.6,0.0],0.1,0.3
Bumblebees: *Bombus hortorum*	296	17.02	99.6,[99.2,0.4],0.1,0.3	99.5,[99.2,0.3],0.3,0.2
Aphids: *Myzus persicae* *	395	69.48	NA	97.1,[94.2,2.9],0.5,2.4
Aphids: *Acyrthosiphon pisum* *	526	126.60	NA	97.6,[94.7,2.9],0.4,2.1
Aphids: *Rhopalosiphum maidis*	326	93.30	97.0,[94.8,2.2],0.7,2.3	98.3,[95.4,2.9],0.5,1.2
Weevil: *Sitophilus oryzae*	770	2.86	97.8,[95.8,2.0],0.7,1.5	98.5,[97.1,1.4],0.3,1.2
Stink bug: *Halyomorpha halys*	1150	0.80	97.4,[96.0,1.4],1.0,1.6	96.7,[95.2,1.5],1.4,1.9
Ladybird: *Cryptolaemus montrouzieri*	988	10.38	97.1,[96.0,1.1],0.6,2.3	97.0,[96.4,0.6],1.1,1.9
Ladybird: *Harmonia axyridis*	417	2.05	92.4,[90.0,2.4],1.2,6.4	91.7,[89.3,2.4],1.4,6.9
Butterfly: *Vanessa cardui* (Ph)	425	14.62	98.9,[98.8,0.1],0.4,0.7	98.9,[98.6,0.3],0.4,0.7
Butterfly: *Vanessa cardui* (Ah)	401	2.75	96.1,[96.0,0.1],0.4,3.5	95.6,[95.4,0.2],0.5,3.9
Moth: *Ectropis grisescens*	785	26.91	96.4,[95.7,0.7],1.2,2.4	95.6,[95.2,0.4],1.9,2.5
Butterfly: *Pieris macdunnoughii*	317	5.20	97.2,[96.3,0.9],0.4,2.4	97.2,[96.5,0.7],0.9,1.9

Completeness assessments with BUSCO v4.1.4 and assembly statistics sourced from the A^{3}Cat [10], or (*) directly from [38]. C=Complete, [S = Complete Single, D = Complete Duplicated], F = Fragmented, M = Missing. Ph = Principal Haplotype, Ah = Alternative Haplotype.

Figure 2



Number of BioProject entries for the 15 most represented insect species. Counts of BioProjects sourced from the United States NCBI (January 2022) show that the classical model species, *Drosophila melanogaster*, is associated with an order of magnitude more registered projects than the other most represented species. Bar colours represent a simplified 'principal research interest/relevance' category for each species. 'Word Clouds' for selected species are built from the collated titles of all their available BioProjects. *Drosophila melanogaster* (Diptera): fruit fly, *Apis mellifera* (Hymenoptera): western honey bee, *Aedes aegypti* (Diptera): yellow fever mosquito, *Bombyx mori* (Lepidoptera): domestic silk moth, *Anopheles gambiae* (Diptera): African malaria mosquito, *Drosophila simulans* (Diptera): fruit fly, *Bemisia tabaci* (Hemiptera): silverleaf whitefly, *Spodoptera frugiperda* (Lepidoptera): fall armyworm, *Aedes albopictus* (Diptera): tiger mosquito, *Bactrocera dorsalis* (Diptera): oriental fruit fly, *Nilaparvata lugens* (Hemiptera): brown planthopper, *Locusta migratoria* (Orthoptera): migratory locust, *Acyrthosiphon pisum* (Hemiptera): pea aphid, *Tribolium castaneum* (Coleoptera): red flour beetle, *Helicoverpa armigera* (Lepidoptera): cotton bollworm.

their eggs on invasive Eurasian mustard plants that are lethal to the larvae.

These examples of emerging models with reference genome assemblies show how technological advances are supporting new models by delivering high-quality data. The reference genomes themselves provide a framework onto which new knowledge can be mapped, from comparative genomic analyses, molecular biology experiments, as well as functional and population genomic datasets (transcriptomics, proteomics, metabolomics, resequencing, etc.). Using the number of NCBI BioProjects as a proxy to gauge the extent of genome-enabled research activities shows how the classical model insect species, *Drosophila melanogaster*, is associated with an order of magnitude more registered projects than the other most represented species (Figure 2). Among the others are well-known species that are economically important, vectors of human diseases, or agricultural pests, all of which have had publicly available draft assemblies for more than 5 years and almost all of which now have published high-quality assembly upgrades, including most recently for the fall armyworm [51], the tiger mosquito [52], the brown planthopper [53], and the red flour beetle [54]. Species representing emerging model systems such as the examples outlined above are expected to similarly build genome-anchored knowledge bases that support and enrich the exploration of the diversity of insect biology and evolution.

Conclusions

New technologies are helping to greatly expand the diversity of insect species for which genome resources are being generated across Insecta [9,10], presenting opportunities to develop new model systems for studying a large variety of biological phenomena. Within-genus sampling is also reaching new levels of resolution, exemplified by the genome assemblies for 101 lines of 93 drosophilid species spanning 14 species groups and 35 subgroups [55]. Nevertheless, challenges such as working with large repeat-rich genomes or very small specimens from which to extract high-molecular-weight DNA mean that achieving reference-quality standards can still be arduous [12]. The active participation of the arthropod genomics community in the development of standards and provision of guidelines and protocols through initiatives coordinating the scaling up of reference genome generation help overcome many of these challenges [22,56,57]. Gene content completeness and other quality assessments during production and of the resulting chromosome-level assemblies will therefore continue to play a key role in establishing genome resources that best support the development of new model systems and advance understanding of insect biology and evolution.

Conflict of interest statement

The authors certify that they have no affiliations with or involvement in any organisation or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Acknowledgements

The authors thank Sagane Dind, Giulia Campli, Livio Ruzzante, and Antonin Thiébaut for providing useful suggestions for improvements and valuable feedback on the manuscript. This work was supported by the Swiss National Science Foundation [grant numbers PP00P3_170664 and PP00P3_202669].

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Ankeny R, Leonelli S: Model Organisms. Cambridge University Press; 2020.
2. Bilder D, Irvine KD: **Taking stock of the *Drosophila* research ecosystem**. *Genetics* 2017, **206**:1227-1236.
3. Adams MD: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**:2185-2195.
4. Matthews BJ, Vosshall LB: **How to turn an organism into a model organism in 10 'easy' steps**. *J Exp Biol* 2020, **223**:jeb218198.
5. Campbell JF, Athanassiou CG, Hagstrum DW, Zhu KY: ***Tribolium castaneum*: a model insect for fundamental and applied research**. *Annu Rev Entomol* 2022, **67**:347-365.
6. Averof M, Sinigaglia C: **Introduction to emerging systems**. *EvoDevo* 2020, **11**:8 (s13227-020-00153-y).
7. Duffy MA, García-Robledo C, Gordon SP, Grant NA, Green DA, Kamath A, Penczykowski RM, Rebolleda-Gómez M, Wale N, Zaman L: **Model systems in ecology, evolution, and behavior: a call for diversity in our model systems and discipline**. *Am Nat* 2021, **198**:53-68.
8. Hotaling S, Kelley JL, Frandsen PB: **Aquatic insects are dramatically underrepresented in genomic research**. *Insects* 2020, **11**:601.
9. Hotaling S, Sproul JS, Heckenhauer J, Powell A, Larracuente AM, Pauls SU, Kelley JL, Frandsen PB: **Long-reads are revolutionizing 20 years of insect genome sequencing**. *Genome Biol Evol* 2021, **13**:evab138, https://doi.org/10.1093/gbe/evab138
10. Feron R, Waterhouse RM: **Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes**. *Gigascience* 2022, **11**:giac006, https://doi.org/10.1093/gigascience/giac006
11. González VL, Devine AM, Trizna M, Mulcahy DG, Barker KB, Coddington JA: **Open access genomic resources for terrestrial arthropods**. *Curr Opin Insect Sci* 2018, **25**:91-98.
12. Li F, Zhao X, Li M, He K, Huang C, Zhou Y, Li Z, Walters JR: **Insect genomes: progress and challenges**. *Insect Mol Biol* 2019, **28**:739-758.
13. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV: **OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs**. *Nucleic Acids Res* 2013, **41**:D358-D365.
14. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA, Ioannidis P, Zdobnov EM: **OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software**. *Nucleic Acids Res* 2015, **43**:D250-D256.
15. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV: **OrthoDB v9.1: cataloging evolutionary and functional annotations for

animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res* 2017, **45**:D744-D749.

16. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs**. *Bioinformatics* 2015, **31**:3210-3212.

17. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM: **BUSCO applications from quality assessments to gene prediction and phylogenomics**. *Mol Biol Evol* 2018, **35**:543-548.

18. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes**. *Mol Biol Evol* 2021, **38**:4647-4654.

19. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM: **OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs**. *Nucleic Acids Res* 2019, **47**:D807-D811.

20. Waterhouse RM, Seppey M, Simão FA, Zdobnov EM: **Using BUSCO to assess insect genomic resources**. In *Insect Genomics*. Edited by Brown SJ, Pfrender ME. Springer; 2019:59-74.

21. Manni M, Berkeley MR, Seppey M, Zdobnov EM: **BUSCO: assessing genomic data quality and beyond**. *Curr Protoc* 2021, **1**:e323.

22. Lawniczak MKN, Durbin R, Flicek P, Lindblad-Toh K, Wei X, Archibald JM, Baker WJ, Belov K, Blaxter ML, Marques Bonet T, *et al*.: **Standards recommendations for the Earth BioGenome Project**. *Proc Natl Acad Sci* 2022, **119**:e2115639118.

23. Feron R, Waterhouse RM: **Supporting data for "Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes"**. *GigaDB* 2022, **11**:giac006, https://doi.org/10.5524/100974

24. Favreau E, Martínez-Ruiz C, Rodrigues Santiago L, Hammond RL, Wurm Y: **Genes and genomic processes underpinning the social lives of ants**. *Curr Opin Insect Sci* 2018, **25**:83-90.

25. Branstetter MG, Childers AK, Cox-Foster D, Hopper KR, Kapheim KM, Toth AL, Worley KC: **Genomes of the Hymenoptera**. *Curr Opin Insect Sci* 2018, **25**:65-75.

26. Panfilio KA, Angelini DR: **By land, air, and sea: hemipteran diversity through the genomic lens**. *Curr Opin Insect Sci* 2018, **25**:106-115.

27. McKenna DD: **Beetle genomes in the 21st century: prospects, progress and priorities**. *Curr Opin Insect Sci* 2018, **25**:76-82.

28. Wiegmann BM, Richards S: **Genomes of Diptera**. *Curr Opin Insect Sci* 2018, **25**:116-124.

29. Ruzzante L, Reijnders MJMF, Waterhouse RM: **Of genes and genomes: mosquito evolution and diversity**. *Trends Parasitol* 2019, **35**:32-51.

30. Triant DA, Cinel SD, Kawahara AY: **Lepidoptera genomes: current knowledge, gaps and future directions**. *Curr Opin Insect Sci* 2018, **25**:99-105.

31. Richards S, Childers A, Childers C: **Editorial overview: Insect genomics: arthropod genomic resources for the 21st century: It only counts if it's in the database!**. *Curr Opin Insect Sci* 2018, **25**:iv-vii.

32. •• Almudi I, Vizueta J, Wyatt CDR, de Mendoza A, Marlétaz F, Firbas PN, Feuda R, Masiero G, Medina P, Alcaina-Caro A, *et al*.: **Genomic adaptations to aquatic and aerial life in mayflies and the origin of insect wings**. *Nat Commun* 2020, **11**:2631.

This study exemplifies the coming together of establishing a new laboratory-tractable system with the generation of genome resources and extensive functional genomics data to support novel biological investigations. The mayfly genome provided the framework to explore patterns of gene expression throughout its aquatic and aerial life cycle and across different organs, and to identify a core set of genes involved in insect wing development.

33. •• Sun C, Huang J, Wang Y, Zhao X, Su L, Thomas GWC, Zhao M, Zhang X, Jungreis I, Kellis M, *et al*.: **Genus-wide characterization of bumblebee genomes provides insights into their evolution and variation in ecological and behavioral traits**. *Mol Biol Evol* 2021, **38**:486-501.

Chromosome-level assemblies generated for 5 of the 17 bumblebee species in this study allowed tracing of the rearrangements that created the unusual 25-chromosome karyotype in social parasites. The high-quality genome resources from sampling species across the genus supported the quantification of genetic and genomic variation across the Bombus phylogeny, where high levels of gene tree discordance are likely driven by incomplete lineage sorting.

34. Christmas MJ, Jones JC, Olsson A, Wallerman O, Bunikis I, Kierczak M, Peona V, Whitley KM, Larva T, Suh A, *et al*.: **Genetic barriers to historical gene flow between cryptic species of alpine bumblebees revealed by comparative population genomics**. *Mol Biol Evol* 2021, **38**:3126-3143.

35. Crowley L, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium: **The genome sequence of the garden bumblebee, *Bombus hortorum* (Linnaeus, 1761)**. *Wellcome Open Res* 2021, **6**:270.

36. Maebe K, Hart AF, Marshall L, Vandamme P, Vereecken NJ, Michez D, Smagghe G: **Bumblebee resilience to climate change, through plastic and adaptive responses**. *Glob Change Biol* 2021, **27**:4223-4237.

37. Li Y, Zhang B, Moran NA: **The aphid X chromosome is a dangerous place for functionally important genes: diverse evolution of hemipteran genomes based on chromosome-level assemblies**. *Mol Biol Evol* 2020, **37**:2357-2368.

38. • Mathers TC, Wouters RHM, Mugford ST, Swarbreck D, van Oosterhout C, Hogenhout SA: **Chromosome-scale genome assemblies of aphids reveal extensively rearranged autosomes and long-term conservation of the X chromosome**. *Mol Biol Evol* 2021, **38**:856-875.

This study highlights the importance of examining genome evolutionary features in different insect clades, especially chromosome rearrangements as they can lead to rapid evolution and speciation. By developing and comparing high-quality genome resources for these aphids, dramatic reorganisations of their autosomes were found to sharply contrast the much more stably maintained gene content of the aphid sex chromosome.

39. • Singh KS, Cordeiro EMG, Troczka BJ, Pym A, Mackisack J, Mathers TC, Duarte A, Legeai F, Robin S, Bielza P, *et al*.: **Global patterns in genomic diversity underpinning the evolution of insecticide resistance in the aphid crop pest *Myzus persicae***. *Commun Biol* 2021, **4**:847.

Using a chromosome-level genome assembly of the aphid, *Myzus persicae*, and 40× short-read sequencing of 127 clones derived from 19 countries, this study shows how a high-quality reference enables comprehensive characterisations of genome-wide patterns of global genetic variation. This system helps understand genomic responses of insects to insecticide exposures and is particularly relevant for the control of agricultural pests.

40. • Parisot N, Vargas-Chávez C, Goubert C, Baa-Puyoulet P, Balmand S, Beranger L, Blanc C, Bonnamour A, Boulesteix M, Burlet N, *et al*.: **The transposable element-rich genome of the cereal pest *Sitophilus oryzae***. *BMC Biol* 2021, **19**:241.

Beyond building resources to guide progress towards novel pest control tools for this beetle, this study establishes the groundwork for developing a new genome-enabled model for investigating the relationships between endosymbionts and their insect hosts. Despite their relatively recent association, the intracellular symbiotic bacterium appears to rely on the rice weevil for several key amino acids and nucleotides.

41. Sparks ME, Bansal R, Benoit JB, Blackburn MB, Chao H, Chen M, Cheng S, Childers C, Dinh H, Doddapaneni HV, *et al*.: **Brown marmorated stink bug, *Halyomorpha halys* (Stål), genome: putative underpinnings of polyphagy, insecticide resistance potential and biology of a top worldwide pest**. *BMC Genom* 2020, **21**:227.

42. Li H-S, Tang X-F, Huang Y-H, Xu Z-Y, Chen M-L, Du X-Y, Qiu B-Y, Chen P-T, Zhang W, Ślipiński A, *et al*.: **Horizontally acquired**

antibacterial genes associated with adaptive radiation of ladybird beetles. *BMC Biol* 2021, **19**:7.

43. •• Chen M, Mei Y, Chen X, Chen X, Xiao D, He K, Li Q, Wu M, Wang S, Zhang F, *et al*.: **A chromosome-level assembly of the harlequin ladybird *Harmonia axyridis* as a genomic resource to study beetle and invasion biology**. *Mol Ecol Resour* 2021, **21**:1318-1332.

In this study, the generation of a chromosome-level genome assembly of this charismatic ladybird additionally identified the X chromosome and Y-linked scaffolds by separately resequencing males and females. These resources support the development of the harlequin as a model for studying invasion biology in insects, and, with more than 200 described colour forms, for investigating the genetics of colour pattern polymorphisms.

44. Boyes D, Crowley L, University of Oxford and Wytham Woods Genome Acquisition Lab, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium: **The genome sequence of the harlequin ladybird, *Harmonia axyridis* (Pallas, 1773)**. *Wellcome Open Res* 2021, **6**:300.

45. Gautier M, Yamaguchi J, Foucaud J, Loiseau A, Ausset A, Facon B, Gschloessl B, Lagnel J, Loire E, Parrinello H, *et al*.: **The genomic basis of color pattern polymorphism in the harlequin ladybird**. *Curr Biol* 2018, **28**:3296-3302 e7.

46. Ando T, Matsuda T, Goto K, Hara K, Ito A, Hirata J, Yatomi J, Kajitani R, Okuno M, Yamaguchi K, *et al*.: **Repeated inversions within a pannier intron drive diversification of intraspecific colour patterns of ladybird beetles**. *Nat Commun* 2018, **9**:3843.

47. Lohse K, Wright C, Talavera G, García-Berro A, Darwin Tree of Life Barcoding collective, Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium: **The genome sequence of the painted lady, Vanessa cardui Linnaeus 1758**. *Wellcome Open Res* 2021, **6**:324.

48. Zhang L, Steward RA, Wheat CW, Reed RD: **High-quality genome assembly and comprehensive transcriptome of the painted lady butterfly *Vanessa cardui***. *Genome Biol Evol* 2021, **13**:evab145.

49. Pan Y, Fang G, Wang Z, Cao Y, Liu Y, Li G, Liu X, Xiao Q, Zhan S: **Chromosome-level genome reference and genome editing of the tea geometrid**. *Mol Ecol Resour* 2021, **21**:2034-2049.

50. Steward RA, Okamura Y, Boggs CL, Vogel H, Wheat CW: **The genome of the margined white butterfly (*Pieris macdunnoughii*): sex chromosome insights and the power of polishing with PoolSeq data**. *Genome Biol Evol* 2021, **13**:evab053.

51. Zhang L, Liu B, Zheng W, Liu C, Zhang D, Zhao S, Li Z, Xu P, Wilson K, Withers A, *et al*.: **Genetic structure and insecticide resistance characteristics of fall armyworm populations invading China**. *Mol Ecol Resour* 2020, **20**:1682-1696.

52. Palatini U, Masri RA, Cosme LV, Koren S, Thibaud-Nissen F, Biedler JK, Krsticevic F, Johnston JS, Halbach R, Crawford JE, *et al*.: **Improved reference genome of the arboviral vector *Aedes albopictus***. *Genome Biol* 2020, **21**:215.

53. Ye Y, Zhang H, Li D, Zhuo J, Shen Y, Hu Q, Zhang C: **Chromosome-level assembly of the brown planthopper genome with a characterized Y chromosome**. *Mol Ecol Resour* 2021, **21**:1287-1298.

54. Herndon N, Shelton J, Gerischer L, Ioannidis P, Ninova M, Dönitz J, Waterhouse RM, Liang C, Damm C, Siemanowski J, *et al*.: **Enhanced genome assembly and a new official gene set for *Tribolium castaneum***. *BMC Genom* 2020, **21**:47.

55. Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, Comeault AA, Peede D, D'Agostino ER, Pelaez J, *et al*.: **Highly contiguous assemblies of 101 drosophilid genomes**. *eLife* 2021, **10**:e66405.

56. i5K Consortium: **The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment**. *J Hered* 2013, **104**:595-600.

57. Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini A, Godoy JA, Höglund J, *et al*.: **The era of reference genomes in conservation genomics**. *Trends Ecol Evol* 2022, **37**:197-202, https://doi.org/10.1016/j.tree.2021.11.008

Chapter 2: A reproducible workflow for Multispecies Whole Genome Alignments

Summary

This thesis chapter summarises the research and development work performed to build a reproducible, portable, and scalable workflow for computing multispecies whole genome alignment (MWGA) datasets. The *mwgaw-align* workflow was developed and tested using publicly available genome assemblies for different groups of arthropods. It employs tools developed by the GenomeBrowser team at the University of California Santa Cruz (UCSC) integrated in a Snakemake workflow which manages all the required data processing steps and enables parallelisation of the compute-intensive pairwise alignments. This chapter describes the implementation and application of the workflow while the technical solutions developed to address specific data processing steps are presented in **Chapter 5**. The *mwgaw-align* workflow was applied to generate MWGA datasets for 22 mosquitoes, 6 tse-tse flies, and 11 bees presented in this chapter to demonstrate the results that can be obtained using the workflow. The outputs of this chapter provide the community with a comprehensive workflow that others can use to build MWGA datasets for their own groups of species in a reproducible and scalable manner. The resulting MWGA datasets also serve as a foundation for the work presented in **Chapter 3** and **Chapter 4**, where the alignments are used as inputs for quantifying sequence conservation and protein-coding potential.

Introduction

The majority of multispecies whole genome alignments (MWGA) published to date were computed following a workflow developed by the GenomeBrowser team at the University of California, Santa Cruz (UCSC). This team played a crucial role in publishing and annotating the first Human genome, and in setting up an ecosystem of utilities which evolved into the modern version of GenomeBrowser. Thanks to this historical role, GenomeBrowser was the original repository for updates to the Human genome, before being superseded by the National Center for Biotechnology Information (NCBI) and later the Genome Reference Consortium (GRC). Naturally, other major genomes released around that time were integrated into GenomeBrowser. The need soon arose to compare these available genomes to identify their shared - and unique

- sequences, mostly to assist with annotating the Human genome (ENCODE Project Consortium, 2012) and to compare gene sequences between species. With this goal in mind, the GenomeBrowser team implemented a workflow to align entire genome assemblies to a reference assembly using BLASTZ (Schwartz et al., 2003), combine these pairwise alignments into a multispecies whole genome alignment using the Threaded-Blockset Aligner (TBA) or roast (Blanchette et al., 2004), estimate sequence conservation across species with phastCons (Siepel et al., 2005), and integrate these results into the Genome Browser. Alignments produced by this workflow are thus intrinsically linked to the GenomeBrowser internal data structure and computational platform, which is, as mentioned previously, focused on the human genome. While this development effort yielded many pieces of software and scripts which are essential in computing whole genome alignments to this day, the entanglement of these tools with GenomeBrowser coupled with a lack of accessible documentation on the workflow to compute MWGAs resulted in a strong dependence on the UCSC GenomeBrowser team to publish alignments. In fact, the vast majority of MWGAs available until recent years were published by or in collaboration with that team (Christmas et al., 2023; Lindblad-Toh et al., 2011; Miller et al., 2007). Others interested in computing alignments, including our group, had to rely on patchy and unorganised information scattered around GenomeBrowser data pages, GenomeBrowser wiki pages, and personal notes; some of these pages have disappeared in the past few years; one could argue that the lack of documentation and convenient workflow is one of the main reasons why so few MWGAs have been computed so far, despite the valuable data they provide to address many scientific questions. At the start of this project, however, there was no viable alternative to the UCSC GenomeBrowser workflow to compute MWGAs. As we planned to compute MWGAs for different arthropod clades, we needed a reliable implementation of that workflow. Such an implementation would benefit the genomics research community both as a way to reproduce past MWGAs and analyses performed on them, and as a tool facilitating computation of new MWGAs outside of the GenomeBrowser ecosystem. Consequently, we decided to invest in the development and publication of a reproducible, documented, and portable implementation of the standard workflow to compute MWGAs. In this chapter, we will describe our implementation of the MWGA workflow and the alignments we computed using it.

Methods

Summary of the *mwgaw-align* workflow

The *mwgaw-align* workflow takes as input one fasta file for each assembly to include in the alignment and a simple tree describing the phylogenetic relationship between the species to which these assemblies belong. These input files are parsed to extract relevant information, and formatted to comply with the requirements of the tools used in the workflow. Among the assemblies provided to the workflow, the user specifies a *reference assembly*, which is the assembly to which all other assemblies are aligned, and the resulting MWGA will use the genomic coordinates of this reference assembly. All pairwise whole genome alignments are merged into the MWGA, which is then ordered and sorted. Each step of this workflow is detailed in the following sections, and a simplified flowchart of the main steps of the workflow is presented in **Figure 2.1** below; for reference, a Directed Acyclic Graph visualisation of the complete workflow for three bacterial assemblies is shown in **Appendix 3**.

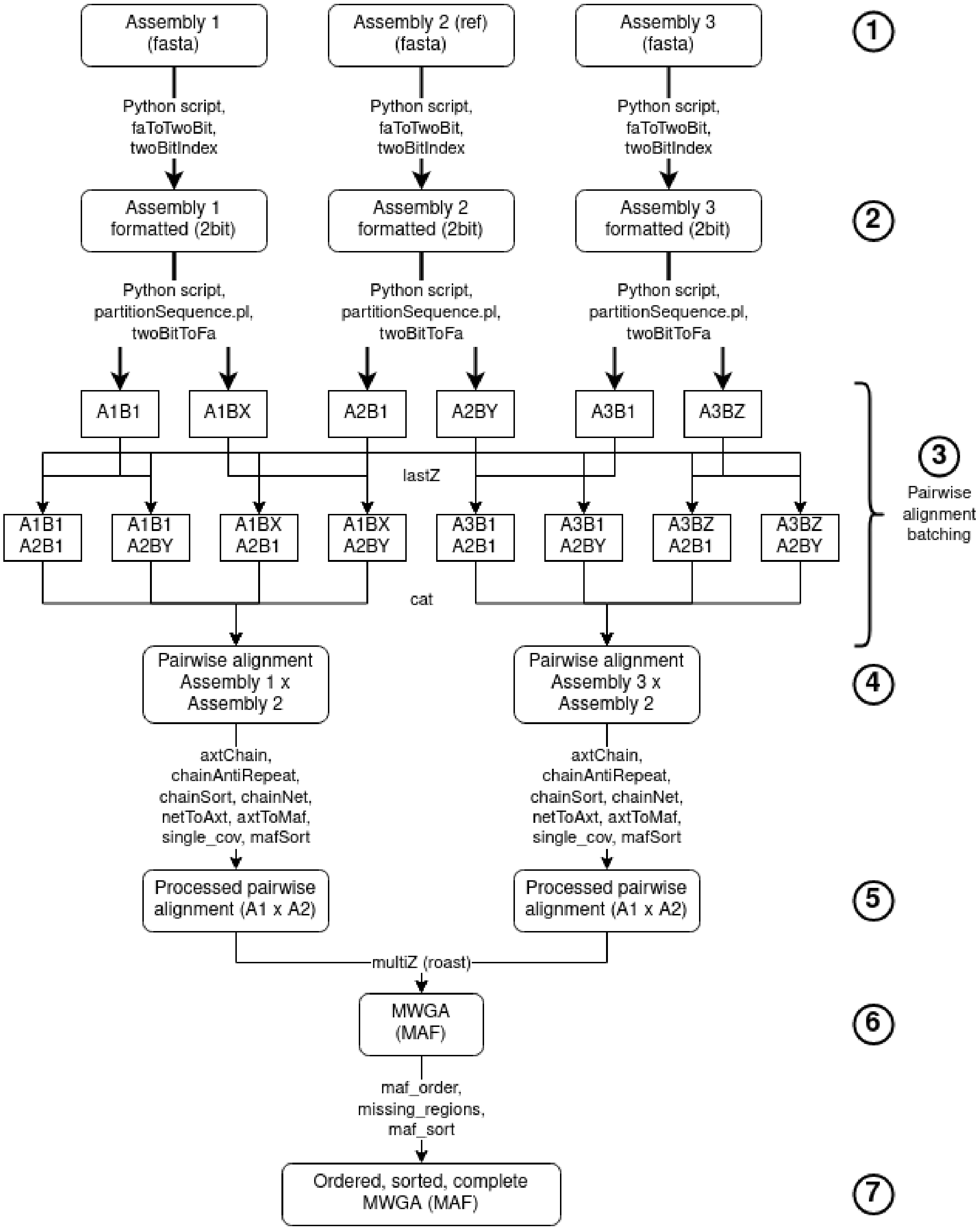


Figure 2.1: Simplified diagram of the *mwgaw-align* workflow. This diagram shows the main steps of the workflow for three assemblies, one reference and two other assemblies (1). Each assembly is first formatted and converted into a binary format (2), before being split into batches. For each non-reference assembly, all batches are individually aligned to all batches from the reference assembly (3), and alignment batches are gathered into a single pairwise alignment for each non-reference assembly (4). These pairwise alignments undergo chaining, sorting, netting, and multiple conversion steps (5) before being merged into a MWGA with the *roast* software from MULTIZ (6). The resulting Multiple Alignment Format (MAF) file is ordered, sorted, and completed to produce the final MAF output of *mwgaw-align* (7).

Processing the assembly and phylogeny inputs

Assemblies to include in the MWGA are provided directly as fasta files by the user in a tabulated file specifying a name for the assembly, which will be used as the assembly's identifier in the final alignment, and the path to the fasta file. Input assemblies are processed with a custom Python script to rename contig headers following the format *assembly:contig*, using only the contig identifier from the original fasta file - *e.g.*, discarding any additional information included in standard headers from NCBI assemblies. This contig header format is required by most tools used in later parts of the workflow. For convenience, and to retain all the information from the original assemblies, a correspondence table between original contig headers and formatted headers is generated during this step. Most pre-existing tools included in the workflow require assemblies in a binary 2bit format; therefore, assemblies with formatted contig headers are converted to this 2bit format using the UCSC utility *faToTwoBit*, and a table of contig lengths is generated with the *twoBitInfo* utility. The resulting formatted assemblies in binary format and the corresponding tables of contig lengths are ready to be used as input for the following alignment step. In parallel, the input Newick phylogenetic tree is converted to the format required by the multiple alignment software MULTIZ using a custom Python script relying on the *ete3* Python library (Huerta-Cepas et al., 2016). The resulting tree is space-separated and without branch lengths, *e.g.* "((holophaga_bacterium holophaga_foetida) geothrix_fermentans)" for three bacterial species.

Partitioning input assemblies to parallelise pairwise alignments

Pairwise alignments are the most CPU-intensive steps of the workflow, at least for MWGAs including a small to medium number of assemblies; for MWGAs including a large number of assemblies, combining pairwise alignments into a MWGA may become the longest step. Regardless, in order to reduce the effective compute time of pairwise alignments, input assemblies are partitioned in batches using a process described in the following paragraph, the size of which is adjusted in the user config file based on the length of assemblies and the desired total number of individual jobs in the workflow. Because the runtime of a pairwise alignment job is proportional to the size of each sequence aligned, a smaller batch size strongly reduces the runtime of pairwise alignment jobs. However, smaller batches also increase the total number of individual jobs multiplicatively, which can dramatically slow down dependency resolution and job selection by the workflow management engine. For these reasons, the ideal batch size should provide a balance between the duration of a single alignment job, with limiting factors like maximum runtime on a computational platform, and total number of jobs, to avoid overloading the workflow management engine and the system to which jobs are submitted.

In practice, a partition file is generated for each assembly using the *partitionSequence.pl* Perl script from UCSC Kent utilities; for the reference assembly, batches are created with a slight overlap in order to minimise the effect of breaking up assembly sequences on the alignment process; overlaps are handled in the downstream merging process. This Perl script generates two types of batches: batches comprising a single sequence, *i.e.* part of a contig which total length is higher than the size of a batch, and batches comprising multiple sequences, *i.e.* multiple entire contigs, each smaller than the size of a batch. All single-sequence batches are exported as entries in an output file by the *partitionSequence.pl* script, while multi-sequence batches are each exported in a separate file in an output folder. All these output files are collected, processed, and combined into a single list of partitions using a custom Python script; this list is required in some of the downstream steps of the workflow.

When aligning sequences from two batches, genomic coordinates in the resulting pairwise alignment are relative to the input sequences. These coordinates need to be converted back to the coordinates in the original assemblies; to do so, a coordinate lift file is created with the *constructLiftFile.pl* script from UCSC Kent utilities, and this file is later used to “lift over” the coordinates from a pairwise alignment.

Computing pairwise whole genome alignments

Each batch of sequence(s) from each non-reference assembly is aligned to each batch of sequence(s) from the reference assembly; to illustrate this process, in an alignment including a reference assembly partitioned into three batches and five non-reference assemblies each partitioned into four batches, the total number of pairwise alignment jobs would be 3 x 5 x 4 = 60 jobs. For a single alignment job, the two batches of sequences to align are first extracted in fasta format from the assembly 2bit files using the UCSC utility *twoBitToFa*, and a table of sequence lengths is generated with the utility *faSize* for each resulting fasta file. The two batches of fasta sequences are then aligned with LASTZ using parameters defined in the config file; default parameter values are based on that from the UCSC 124 insects conservation tracks (https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=dm6&g=cons124way). By default, LASTZ generates alignments in *axt* format, which stores the entire sequences in each alignment block; these files are directly converted to *psl* format, which is more compact and readable by all downstream tools, using the *axtToPsl* utility. When sequences are extracted from the assembly files, their original genomic coordinates are lost, and the alignment generated by LASTZ uses coordinates relative to the actual sequences aligned. Therefore, genomic coordinates in the alignment need to be "lifted", *i.e.* converted back to the coordinates in the original assembly file, using the *liftUp* utility and the coordinate lift file mentioned above. The final result of a pairwise alignment of a batch of sequences from the reference assembly and a batch of sequences from one of the non-reference assembly is a compact *psl* format with genomic coordinates from the original assemblies.

Then, for each non-reference assembly, all pairwise alignments of a batch of sequences to a batch from the reference assembly are concatenated into a final alignment in *psl* format with the *cat* unix utility and directly compressed with *gzip* to reduce disk space. Although LASTZ can merge local alignments into ungapped blocks called chains, this process is disabled in our default settings. Instead, chaining is performed with the *axtChain* UCSC utility. Chains obtained with this approach are defined as a succession of non-overlapping gapless alignment blocks separated by gaps on either one or on both sequences, similar to the final output of LASTZ. Computing chains with *axtChain* instead of relying on LASTZ's output allows to filter out sequences with very strong bias towards one or two nucleotides (degenerated sequences) and sequences that are composed mainly of repeated elements that were masked in the input assembly using tools like RepeatMasker (repeated sequences). Degenerated and repeated sequences are filtered out using the *chainAntiRepeat* utility and the resulting chains are extended into nets, which are an organised collection of chains sorted by score so that the reference sequence is only covered once. To obtain nets, chains are first sorted with *chainSort* and cleaned with *chainPreNet* to filter out small chains entirely contained within higher-scoring chains; the final net output is obtained with *chainNet*. The resulting alignments are then converted back to *axt* format with *netToAxt*, sorted by reference sequence position with *axtSort*, and finally converted to Multiple Alignment Format (MAF) with *axtToMaf*. The multiple format conversions are required because the tools used in the steps described above have different input format requirements. The resulting MAF files are filtered with the *single_cov2* software from the MULTIZ package to ensure that both reference and non-reference sequences are covered only once in the alignment, and the final MAF output is ordered to ensure that sequences always appear in the same order within an alignment block, and finally sorted by genomic coordinates on the reference sequence. The final result of a pairwise alignment of a non-reference assembly to the reference assembly is a sorted, ordered, and single-coverage MAF file.

Merging pairwise alignments into an MWGA

Pairwise alignments are integrated into a MWGA with the *roast* software from MULTIZ. This software requires that all MAF pairwise alignments and all fasta assembly files are located in the same directory. Therefore, the first step of the MWGA process is to create symbolic links for all pairwise alignments and all assembly files in a temporary directory, and *roast* is then run within this directory. The result of running *roast* is a MAF file with merged alignment blocks from all the input pairwise alignment files. This MAF file is then ordered so that assemblies appear in the same order in all alignment blocks using the *maf_order* tool included in MULTIZ. Regions from the reference assemblies that were not aligned to any non-reference assembly are added to the MWGA with our *missing_regions* software from mwga-utils (https://github.com/RomainFeron/mwga-utils, see **Chapter 5** for details), and the resulting MAF file is sorted by position on the reference sequence using the *maf_sort* utility included in MULTIZ. The final output from this step - and from the workflow - is a multispecies whole genome alignment file in MAF format that contains the entire reference sequence, is sorted by genomic coordinates on the reference sequence, and in which assemblies are ordered consistently within each alignment block.

Reproducible, portable, and scalable implementation of the workflow

The workflow is implemented using the Snakemake workflow management engine (Köster & Rahmann, 2012). Snakemake provides both a language to describe a workflow and an engine to execute the workflow for specific inputs and parameter values. This engine can be executed on a local machine, in the cloud, and can also distribute tasks to a load manager on a computational platform (*e.g.* SLURM, the load manager used by the UNIL computational platform Curnagl). The Snakemake language is based on the GNU Make paradigm (Stallman et al., 2004), but implemented as an extension of Python, thus inheriting the easily readable syntax and convenient libraries included in this programming language. Furthermore, in a Snakemake workflow following official guidelines, users interact with the workflow via a well-documented config file using the easily readable YAML syntax. These features make Snakemake workflows easy to use and to understand, which helps detecting potential implementation errors when reviewed by peers. Altogether, Snakemake strongly facilitates the implementation of entirely reproducible, verifiable, and portable workflows, a requirement for modern bioinformatics analyses.

Snakemake can use the Conda environment and package manager to deploy specific versions of the software used in a workflow; all pieces of software used to compute MWGAs and perform downstream analyses were either already available in the Bioconda channel (Grüning et al., 2018) or were packaged by us if missing (a detailed list of packaged software is provided in **Chapter 5**), making the workflow entirely portable on all platforms supported by Conda. The workflow was implemented following Snakemake's official guidelines for workflow organisation and implementation, making it easy to use by experienced Snakemake users and efficient to review by experienced Snakemake developers. The config file, repository, and implementation are thoroughly documented to promote transparency and encourage improvements by the community. Finally, we developed a batching system to optimise performance-critical steps, which will help to compute MWGAs including a large number of high-quality assemblies.

Results

To test our *mwgaw-align* workflow described in the previous section, we computed MWGAs for multiple arthropod clades which were of particular interest to us or to our collaborators. In this section, we describe the data used to compute these MWGAs as well as the resulting alignments. We also report runtime and memory usage for different steps of the workflow to illustrate our efforts in optimising performance-heavy steps of the alignment process.

MWGA of 22 mosquito species

The first alignment computed using our workflow was an alignment of 22 mosquito species, using *Anopheles gambiae* as the reference assembly. As the vectors of some of the viruses and parasites deadliest to humans, mosquitoes are the target of considerable research interest both in the medical field and in evolutionary biology. We computed this alignment with the goal of improving existing annotations of mosquito assemblies to assist research efforts, identifying putative functional elements involved in interactions with pathogens, and exploring evolutionary patterns of sequence conservation on mosquito coding sequences; this latter question is the focus of **Chapter 4** (**Applying the MWGA workflows to explore functional constraints in immunity genes**). Pathogen-vectoring mosquitoes are mainly represented by three genera in the Culicidae family: *Aedes*, which includes *Aedes aegypti*, vector of Dengue fever, Yellow fever, Zika, and Chikungunya among many other viruses; *Culex*, which species transmit West Nile virus, several encephalitis, and several nematodes; and *Anopheles*, the genus of *Anopheles gambiae*, vector of human malaria and model species for mosquito research. We included assemblies from these three genera in our mosquito genome alignment, with the vast majority of assemblies belonging to *Anopheles* species, reflecting both the higher diversity of species and sampling bias in genome assemblies in this genus compared to *Culex* and *Aedes*. The full list of species included in our alignment, the assembly used, assembly size, as well as contiguity and completeness are presented in **Table 2.1 below**.

Species	**GenBank Accession**	**Length**	**Scaffold number**	**Scaffold N50**	**BUSCO Score**
Aedes aegypti	GCA_002204515.1	1,278,715,314	2,309	409,777,670	93.6
Aedes albopictus	GCA_001444175.2	1,923,476,627	154,782	201,017	75
Anopheles albimanus	GCA_000349125.2	173,339,239	236	9,735,467	96.9
Anopheles arabiensis	GCA_000349185.1	246,567,867	1,214	5,604,218	97.2
Anopheles atroparvus	GCA_000473505.1	224,290,125	1,371	9,206,694	96.9
Anopheles christyi	GCA_000349165.1	172,658,580	30,369	9,057	96.2
Anopheles coluzzii	GCA_000150765.1	224,417,174	10,521	4,437,438	93.5
Anopheles culicifacies	GCA_000473375.1	202,998,806	16,162	22,320	96.2
Anopheles darlingi	GCA_000211455.3	136,935,538	2,220	115,072	96.3
Anopheles dirus	GCA_000349145.1	216,307,690	1,266	6,906,475	97.6
Anopheles epiroticus	GCA_000349105.1	223,486,714	2,673	366,526	97.5
Anopheles farauti	GCA_000473445.2	183,103,254	310	12,895,223	97.6
Anopheles funestus	GCA_000349085.1	225,223,604	1,392	671,960	97.4
Anopheles gambiae	GCA_000005575.1	265,011,681	8,144	12,309,988	96.4
Anopheles maculatus	GCA_000473185.1	141,894,015	47,797	3,841	70.3
Anopheles melas	GCA_000473525.2	224,162,116	20,229	18,103	92.6
Anopheles merus	GCA_000473845.2	288,048,996	2,027	1,489,982	97.6
Anopheles minimus	GCA_000349025.1	201,793,324	678	10,313,149	97.9
Anopheles quadriannulatus	GCA_000349065.1	283,828,998	2,823	1,641,272	97.4
Anopheles sinensis	GCA_000472065.2	375,763,635	10,448	579,086	94.1
Anopheles stephensi	GCA_000349045.1	225,369,006	1,110	837,295	97.3
Culex quinquefasciatus	GCA_000209185.1	579,042,118	3,171	486,756	90.6

Table 2.1: species, assembly accession, assembly length in base pairs, number of scaffolds, scaffold N50, and BUSCO score for all assemblies included in the 22 mosquitoes alignment. The assembly for *Anopheles gambiae* (GCA_000005575.1) is used as reference for the alignment, as it is by far the best annotated mosquito assembly. Other assemblies belong to members of the *Anopheles* genus, except for *Culex quinquefasciatus*, *Aedes aegypti*, and *Aedes albopictus* which represent the *Culex* and *Aedes* genera, respectively.

The size of assemblies ranged from ~142 Mbp to ~376 Mbp for *Anopheles* species; these values are typical for arthropod genome assemblies, which have a median size of ~360 Mbp and an average size of ~250 Mbp (N=5627, available Arthropod assemblies obtained from A^{3}Cat release 2024-02-01). *Culex* and *Aedes* assemblies are bigger, from ~579 Mbp for *Culex quinquefasciatus* to ~2Gbp for *Aedes albopictus*. Assembly contiguity, one of the variables used to estimate assembly quality, is estimated both from the total number of scaffolds in the assembly and from the scaffold N50 value, which is the length of the scaffold so that the cumulative length of all scaffolds longer than that scaffold adds up to 50% of the assembly span; therefore, a high N50 value indicates that the assembly is highly contiguous. The contiguity of assemblies included in the mosquitoes alignment varied greatly: scaffold N50 ranged from >10 Mbp for chromosome-level assemblies like the reference *Anopheles gambiae*, to an extremely low 3,841 bp for *Anopheles maculatus*. It is worth noting that assembly contiguity does not strongly affect the results of the alignment process, because assemblies are split into small regions (seeds) to generate small local alignments in the first step of the alignment process. However, including contiguous assemblies in the alignment facilitates the interpretation of downstream results, for instance comparison of large genomic regions between species.

A better indicator of assembly quality for MWGAs is assembly completeness, which is estimated with BUSCO; here, we report the percentage of single-copy, complete BUSCOs found in each assembly using the *Arthropoda* dataset. The principles underlying BUSCO and interpretation of BUSCO results are detailed in Chapter 1; here, we observe that 20 out of 22 assemblies had a BUSCO score higher than 90%, 15 of which were higher than 95%, indicating an overall high completeness for these assemblies. Lower BUSCO scores were usually - but not always - correlated with low contiguity, with the highly fragmented *Anopheles maculatus* assembly only reaching 70% single-copy complete BUSCOs. Assemblies of *Anopheles maculatus* and *Aedes albopictus* were retained for the alignment despite their relatively lower quality to maintain the evolutionary resolution of the alignment (for instance, *Aedes albopictus* was one of the only two available *Aedes* assemblies), with the goal of upgrading the alignment once high quality assemblies were released for these species. The maximum divergence time between species included in the alignment was ~150 My, which corresponds to the divergence between the Culicinae and *Anopheles* mosquitoes; divergence time within *Anopheles*, to which most species in the alignment dataset belong, is ~100 My (Neafsey et al., 2015).

Pairwise alignments were generated using lastZ parameter values from the UCSC 124 insect alignment (see workflow description above): HoxD55 scoring matrix, chaining disabled (C=0), dynamic target masking disabled (M=0), a gap open penalty of 400 (O=400), a gap extension penalty of 30 (E=30), a minimal score threshold to perform more sensitive alignment for gapped block of 2,000 (H=2000), a minimal HSP score for the x-drop extension method of 2,200 (K=2200), a minimal score to retain gapped extension blocks of 4,000 (L=4000), a maximum drop-off score for gapped extension of 3,400 (Y=3400), and seeds set to a 19 bp word with at least 12 matching positions, allowing one match to be a transition instead (T=1).

The cumulative runtime of *mwgaw-align* to generate this mosquito MWGA was 10 days, 16 hours, and 13 minutes (total wall time); the longest step was generating the MWGA from pairwise alignments, which lasted for 11 hours and 50 minutes. In comparison, the user runtime - the actual time between the start and the end of the workflow - was one day, 12 hours, and 58 minutes, illustrating the time gains from parallelising pairwise alignment steps; runtime and peak memory usage for the pairwise alignment steps for each non-reference assembly are detailed in **Table 2.2** below. The highest memory usage also occurred during the computation of the MWGA from the pairwise alignments, with a peak memory usage of ~24 Gb.

Species	**Runtime (H:M:S)**	**Memory (Mb)**
Aedes aegypti	18:05:00	2,681
Aedes albopictus	36:23:46	1,752
Anopheles albimanus	12:02:57	1,879
Anopheles arabiensis	14:57:15	1,884
Anopheles atroparvus	14:32:18	1,981
Anopheles christyi	17:06:43	1,750
Anopheles coluzzii	15:50:37	1,867
Anopheles culicifacies	16:32:56	1,744
Anopheles darlingi	8:48:04	1,742
Anopheles dirus	14:08:45	1,857
Anopheles epiroticus	14:52:47	1,776
Anopheles farauti	12:58:14	1,892
Anopheles funestus	14:18:46	1,764
Anopheles maculatus	12:06:20	1,733
Anopheles melas	18:52:54	1,796
Anopheles merus	15:04:03	1,847
Anopheles minimus	13:26:58	1,883
Anopheles quadriannulatus	14:51:16	1,835
Anopheles sinensis	14:51:58	1,757
Anopheles stephensi	13:49:10	1,763
Culex quinquefasciatus	14:02:04	1,745

Table 2.2: total runtime in hours, minutes, and seconds, and peak memory usage of the pairwise alignment step - including chaining, netting, and file conversion - for each assembly on the reference assembly of *Anopheles gambiae*.

The final MWGA for these 22 mosquito species includes 2,545,524,446 aligned bases across all assemblies. The reference assembly *Anopheles gambiae* is contained entirely in the MWGA, with regions aligned to no other assemblies included in the alignment using *mwga-utils* (see methods). For non-reference assemblies, the fraction of aligned bases ranged from 1.26% for *Aedes albopictus* to 95.55% for *Anopheles coluzzii*, with the full results presented in **Table 2.3** below. As expected, assemblies of species evolutionarily closest to *Anopheles gambiae* generally had the highest proportion of aligned bases, *e.g.* species from the *gambiae* complex:

Anopheles arabiensis (~88%), *Anopheles melas* (~91%), *Anopheles merus* (~73%), A*nopheles quadriannulatus* (~75%), and *Anopheles coluzzii* (~96%). In contrast, the assemblies of species most distant to the reference species - namely, species in the *Culicinae* subfamily - had the lowest proportion of aligned bases: *Aedes albopictus* (~1%), *Aedes aegypti* (~2%), *and Culex quinquefasciatus* (~4%). These low numbers are exacerbated by the fact that assemblies of species in the *Culicinae* subfamily are 5 to 8 times bigger than that of the reference species *Anopheles gambiae*, and this size difference is not a result of a recent genome duplication but rather due to the proliferation of transposable elements. The final alignment is summarised in **Figure 2.2** showing assembly size, evolutionary distance to the reference, pairwise alignment runtime, and proportion of aligned bases for all the assemblies included in the alignment.

Assembly	Size	Scaffold N50	BUSCO Score	BP Aligned	BP Aligned (%)
Aedes aegypti	1,278,715,314	409,777,670	93.6	25,145,181	1.97
Aedes albopictus	1,923,476,627	201,017	75	24,194,569	1.26
Anopheles albimanus	173,339,239	9,735,467	96.9	48,511,410	27.99
Anopheles arabiensis	246,567,867	5,604,218	97.2	216,312,367	87.73
Anopheles atroparvus	224,290,125	9,206,694	96.9	81,005,961	36.12
Anopheles christyi	172,658,580	9,057	96.2	156,655,212	90.73
Anopheles coluzzii	224,417,174	4,437,438	93.5	214,426,655	95.55
Anopheles culicifacies	202,998,806	22,320	96.2	128,699,238	63.4
Anopheles darlingi	136,935,538	115,072	96.3	43,315,036	31.63
Anopheles dirus	216,307,690	6,906,475	97.6	115,358,385	53.33
Anopheles epiroticus	223,486,714	366,526	97.5	160,544,258	71.84
Anopheles farauti	183,103,254	12,895,223	97.6	108,854,839	59.45
Anopheles funestus	225,223,604	671,960	97.4	138,504,685	61.5
Anopheles maculatus	141,894,015	3,841	70.3	94,426,810	66.55
Anopheles melas	224,162,116	18,103	92.6	204,175,974	91.08
Anopheles merus	288,048,996	1,489,982	97.6	212,345,680	73.72
Anopheles minimus	201,793,324	10,313,149	97.9	134,181,073	66.49
Anopheles quadriannulatus	283,828,998	1,641,272	97.4	212,802,296	74.98
Anopheles sinensis	375,763,635	579,086	94.1	69,887,087	18.6
Anopheles stephensi	225,369,006	837,295	97.3	131,499,269	58.35
Culex quinquefasciatus	579,042,118	486,756	90.6	24,678,461	4.26

Table 2.3: size, scaffold N50, BUSCO score (as defined in **Table 2.1** above), number of bases aligned in the 22 mosquitoes MWGA and the corresponding percentage of the assembly for each non-reference assembly included in the MWGA.

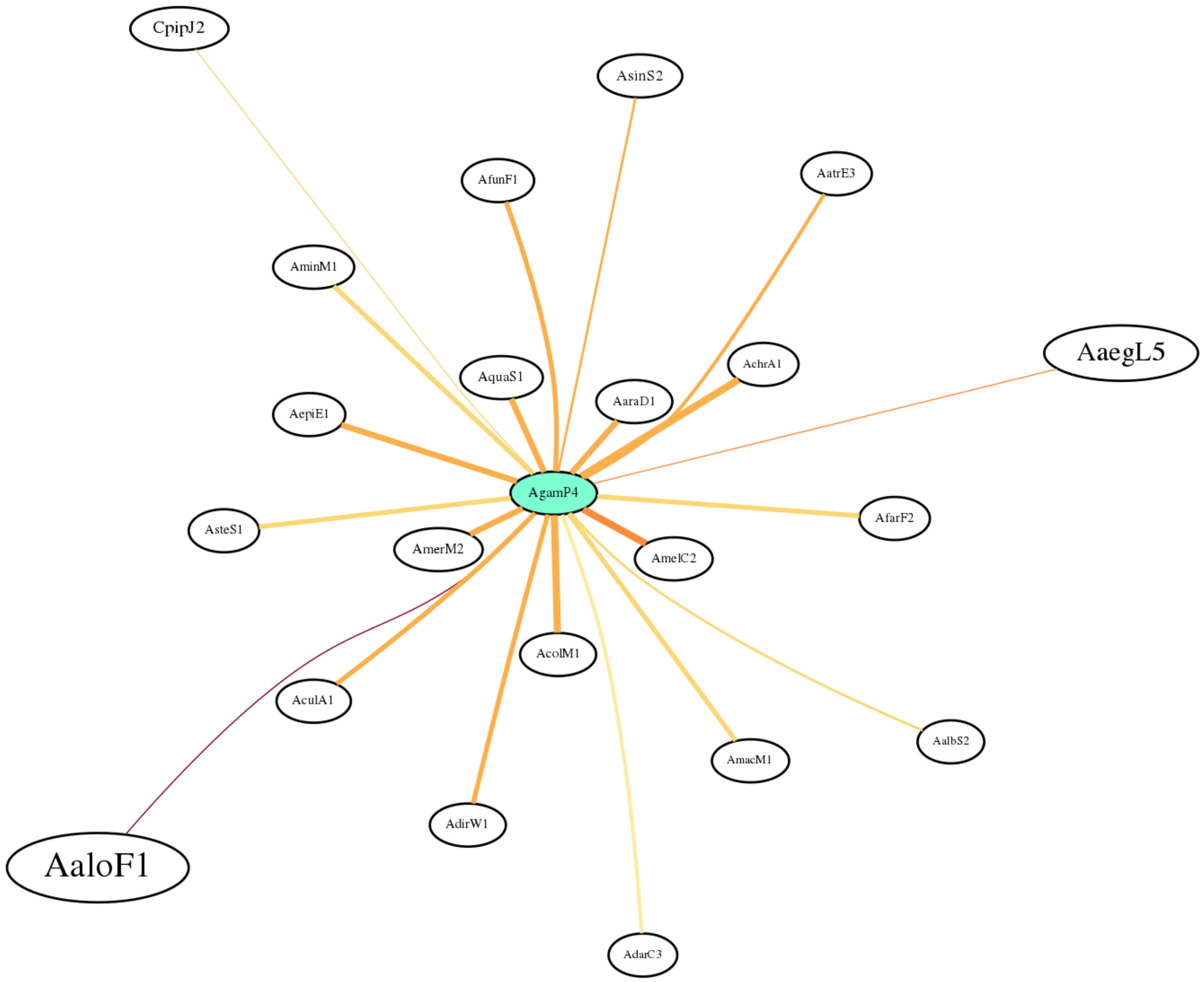


Figure 2.2: graph representation of the 22 mosquitoes Multispecies Whole Genome Alignment (MWGA). Each ellipse represents one assembly included in the alignment, with the reference assembly of *Anopheles gambiae* in the centre, highlighted in green. The size of each ellipse is proportional to the size of the corresponding assembly in bp. The physical distance between an ellipse and the central *Anopheles gambiae* ellipse is proportional to the evolutionary distance between the corresponding species. The thickness of the connection between two ellipses represents the proportion of bases from the non-reference assemblies that were aligned to the reference assembly in the MWGA, and the colour of the connection indicates the total runtime of the pairwise alignment step between the two corresponding assemblies. Expectedly, this visualisation shows that species evolutionarily closer to the reference, *i.e.* physically closer to the centre of the graph, had a higher proportion of aligned bases in the final MWGA than species more evolutionarily distant. However, there was no striking correlation between total runtime and any other variable included in this graph: for instance, both *Aedes albopictus*, a large and distant assembly, and *Anopheles melas*, a small and close assembly, had a high total runtime.

As the first MWGA computed using our workflow, this mosquito alignment was used to test and finish implementing *mwgaw-align* and was also the main test dataset to develop the *mwgaw-analyses* workflow described in **Chapter 3**. As a result, this was the first dataset for which we computed all sequence conservation metrics, and we included these metrics in a GenomeBrowser track hub that was hosted on the GenomeBrowser instance of the former VectorBase website (which is now part of the larger VEuPathDB portal); this track hub is described in the results section of **Chapter 3**. In addition, we used the sequence conservation results in an analysis of evolutionary genomic diversity versus population-level genomic diversity, which is the focus of **Chapter 4**.

MWGA of 6 *Glossina* species (Tse-tse flies)

Another infamous parasite-vectoring clade are tse-tse flies, comprising all species in the genus *Glossina*, which transmits trypanosomes responsible for trypanosomiasis (sleeping sickness). Both mosquitoes and tse-tse flies belong to the order *Diptera* and diverged around 240 Mya, and thus comparing patterns of sequence conservation across their genomes could provide insights on evolutionary adaptations to harbouring pathogens. In a first step towards enabling such comparisons, we computed an MWGA including all available *Glossina* assemblies, which are summarised in the **Table 2.4** below.

Species	**GenBank Accession**	**Length**	**Number of scaffolds**	**Scaffold N50**	**BUSCO Score**
Glossina austeni	GCA_000688735.1	370,264,922	2,205	812,585	97.5
Glossina brevipalpis	GCA_000671755.1	315,360,362	1,651	1,209,507	97.2
Glossina fuscipes	GCA_000671735.1	374,774,708	2,395	561,190	97.2
Glossina morsitans	GCA_001077435.1	363,107,242	24,071	49,769	96.6
Glossina pallidipes	GCA_000688715.1	357,332,231	1,726	1,038,751	97.5
Glossina palpalis	GCA_000818775.1	380,104,241	3,926	575,037	94.4

Table 2.4: species, assembly accession, assembly length in base pairs, number of scaffolds, scaffold N50, and BUSCO score for all assemblies included in the 6 *Glossina* alignment. The assembly for *Glossina morsitans* was used as reference because it is a major vector species and the model species for tse-tse flies.

The sizes of *Glossina* assemblies included in the alignment are homogeneous, ranging from ~315 Mbp to ~380 Mbp, similar to the average arthropod assembly size (median ~360 Mbp, average ~250 Mbp, N=5627). Assembly contiguity, estimated by the number of scaffolds and scaffold N50, is also homogeneous across assemblies with the exception of *Glossina morsitans*, whose assembly is more fragmented than that of other species (~10x higher number of scaffolds, ~10-20x lower N50); this discrepancy originates from the fact that the genome of *Glossina morsitans* was sequenced and assembled before and using older technologies than those of other *Glossina* species. Assembly completeness, estimated using the percentage of complete, single-copy BUSCOs from the *Arthropoda* dataset, is consistently high (>95 except for *Glossina palpalis* with a score of 94.4) across all included assemblies. While contiguity is not high according to modern standards (Feron & Waterhouse, 2022a), with none of the assemblies reaching chromosome level, it is high enough to yield large alignment blocks; furthermore, assembly completeness is more important than contiguity for MWGAs, and therefore these *Glossina* assemblies are of good quality for whole genome alignments.

The final MWGA for these 6 *Glossina* species includes 1,885,615,996 aligned bases across all assemblies. Similar to the mosquito alignment, the reference assembly *Glossina morsitans* is contained entirely in the MWGA, with regions aligned to no other assemblies included in the alignment using *mwga-utils* (see methods). For non-reference assemblies, the fraction of aligned bases ranged from 69.81% for *Glossina brevipalpis* to 95.34% for *Glossina pallidipes*, with the full results presented in **Table 2.5** below. As expected, and as we observed in the mosquito alignment, the proportion of aligned bases in the MWGA was directly related to the evolutionary distance between the assembly's species and the reference *Glossina morsitans*: *Glossina pallidipes* is the closest, followed by *Glossina austeni*, then *Glossina fuscipes* and *Glossina palpalis*, and finally *Glossina brevipalpis*. The final alignment is summarised in **Figure 2.3** showing assembly size, evolutionary distance to the reference, pairwise alignment runtime, and proportion of aligned bases for all the assemblies included in the alignment.

Assembly	**Size**	**Scaffold N50**	**BUSCO Score**	**BP Aligned**	**BP Aligned (%)**
Glossina austeni	370,264,922	812,585	97.5	333,704,024	90.13
Glossina brevipalpis	315,360,362	1,209,507	97.2	220,160,746	69.81
Glossina fuscipes	374,774,708	561,190	97.2	316,358,767	84.41
Glossina pallidipes	357,332,231	1,038,751	97.5	340,680,703	95.34
Glossina palpalis	380,104,241	575,037	94.4	308,515,900	81.17

Table 2.5: size, scaffold N50, BUSCO score (as defined in **Table 2.4** above), number of bases aligned in the 6 Glossina MWGA and the corresponding % of the assembly for each non-reference assembly included in the MWGA.

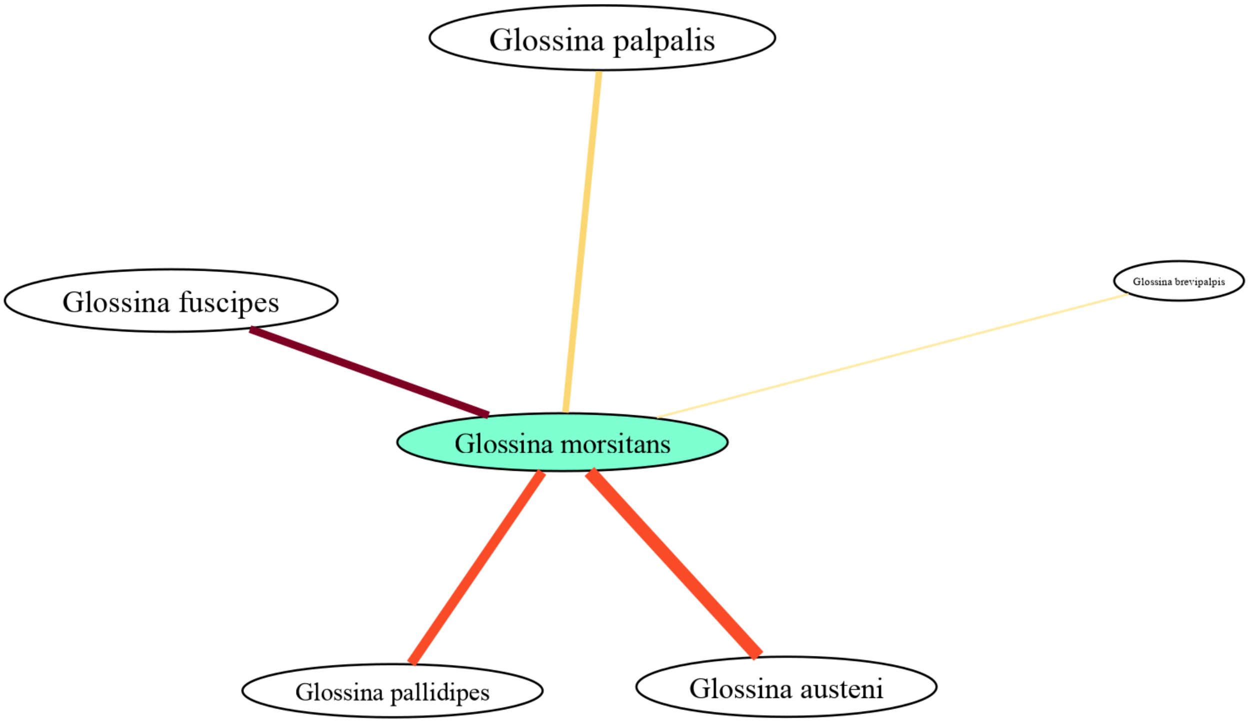


Figure 2.3: graph representation of the 6 *Glossina* Multispecies Whole Genome Alignment (MWGA). Each ellipse represents one assembly included in the alignment, with the reference assembly of *Glossina morsitans* in the centre, highlighted in green. The size of each ellipse is correlated to the size of the corresponding assembly in bp. The physical distance between an ellipse and the central *Glossina morsitans* ellipse is proportional to the phylogenetic distance between the corresponding species. The thickness of the connection between two ellipses represents the proportion of bases from the non-reference assemblies that were aligned to the reference assembly in the MWGA, and the colour of the connection indicates the total runtime of the pairwise alignment step between the two corresponding assemblies (yellow meaning faster, and red meaning slower).

We computed the first version of the *Glossina* MWGA shortly after the mosquito MWGA, as the *Glossina* alignment was of interest to collaborators in Kenya working on the characterisation of immune genes in this clade. We computed an updated version of the MWGA later in this project to incorporate updates to *Glossina* assemblies as well as technical improvements in our workflow, also provided to our Kenyan collaborators.

MWGA of 11 bee species

As part of a collaborative research project centred on a new genome assembly for the Hymenoptera species *Tetragonula carbonaria* (an Australian stingless bee), our group computed a set of evolutionary metrics to investigate the evolutionary trajectories of genes both at the *Hymenoptera* (Order) and at the *Apidae* (Family) levels. Among these metrics were evolutionary sequence conservation, which was estimated at the gene level using an MWGA we computed for 11 bee species. Assemblies included in the alignment are described in the **Table 2.6** below:

Species	**GenBank Accession**	**Length**	**Number of scaffolds**	**Scaffold N50**	**BUSCO Score**
Apis mellifera	GCA_000002195.1	250,270,657	5,644	997,192	97.9
Bombus impatiens	GCA_000188095.4	246,856,484	5,460	1,399,493	99.7
Ceratina calcarata	GCA_001652005.1	199,936,949	50,568	632,424	96.2
Dufourea novaeangliae	GCA_001272555.1	279,505,983	4,178	2,549,405	99.7
Eufriesea mexicana	GCA_001483705.2	595,427,025	187,373	352,794	97.8
Euglossa dilemma	GCA_002201625.1	588,199,719	22,698	143,590	92.4
Habropoda laboriosa	GCA_001263275.1	296,954,667	27,566	1,784,116	98.8
Lasioglossum albipes	GCA_000346575.1	336,521,443	12,599	628,061	97.3
Megachile rotundata	GCA_000220905.1	272,660,569	6,266	1,699,680	99.5
Melipona quadrifasciata	GCA_001276565.1	256,302,695	2,866	1,864,352	99.2
Tetragonula carbonaria	No accession	288,375,904	1,534	16,281,332	-

Table 2.6: species, assembly accession, assembly length in base pairs, number of scaffolds, scaffold N50, and BUSCO score for all assemblies included in the 11 bees alignment. The assembly for *Tetragonula carbonaria* was used as reference as it was the species of interest for this collaborative project; this assembly has not yet been released and thus does not have an accession number.

The size of assemblies selected for the alignment ranged from ~200 Mb to ~600Mb, similar to the average arthropod assembly size (median ~360 Mbp, average ~250 Mbp, N=5627). These assemblies were highly contiguous, as illustrated by their high Scaffold N50, despite not being chromosome-level; completeness was also high (Arthropoda complete single-copy BUSCOs >95%) with the exception of *Euglossa dilemma*, with a lower BUSCO score of 92.4%. Overall, assemblies included in this Bee MWGA were of high quality.

The resulting MWGA for these 11 bee species comprises 1,934,378,481 aligned bases across all assemblies. The proportion of aligned bases ranged from 29.51% for *Euglossa dilemma* to 97.67% for *Melipona quadrifasciata*, with the full results presented in **Table 2.7** below. As observed in the previously described alignments, the proportion of aligned based in an assembly was related to the evolutionary distance between the focal species and *Tetragonula carbonaria*. However, the two species from the *Euglossini* tribe, *Euglossa dilemma* and *Eufriesea mexicana*, appeared as outliers in this trend: they had the lowest proportions of aligned bases (29.51% and 31.47% respectively) while being closer to *Tetragonula carbonaria* than most other species. This observation may be explained by the larger assembly size for these two species, almost double the average size of other assemblies in the alignment: the absolute number of bases aligned for both assemblies follows the trend with respect to the evolutionary distance, but the larger size decreases the proportion of aligned bases. This larger size is likely the result of an accumulation of transposable elements (Brand et al., 2017). The final alignment is summarised in **Figure 2.4** showing assembly size, evolutionary distance to the reference, pairwise alignment runtime, and proportion of aligned bases for all the assemblies included in the alignment.

Assembly	**Size**	**Scaffold N50**	**BUSCO Score**	**BP Aligned**	**BP Aligned (%)**
Apis mellifera	250,270,657	997,192	97.9	158,118,896	63.18
Bombus impatiens	246,856,484	1,399,493	99.7	205,446,432	83.23
Ceratina calcarata	199,936,949	632,424	96.2	111,292,323	55.66
Dufourea novaeangliae	279,505,983	2,549,405	99.7	136,134,911	48.71
Eufriesea mexicana	595,427,025	352,794	97.8	187,358,233	31.47
Euglossa dilemma	588,199,719	143,590	92.4	173,587,336	29.51
Habropoda laboriosa	296,954,667	1,784,116	98.8	168,404,995	56.71
Lasioglossum albipes	336,521,443	628,061	97.3	120,424,779	35.79
Megachile rotundata	272,660,569	1,699,680	99.5	134,896,449	49.47
Melipona quadrifasciata	256,302,695	1,864,352	99.2	250,339,480	97.67

Table 2.7: size, scaffold N50, BUSCO score (as defined in **Table 2.6** above), number of bases aligned in the 11 bees MWGA and the corresponding percentage of the assembly for each non-reference assembly included in the MWGA.

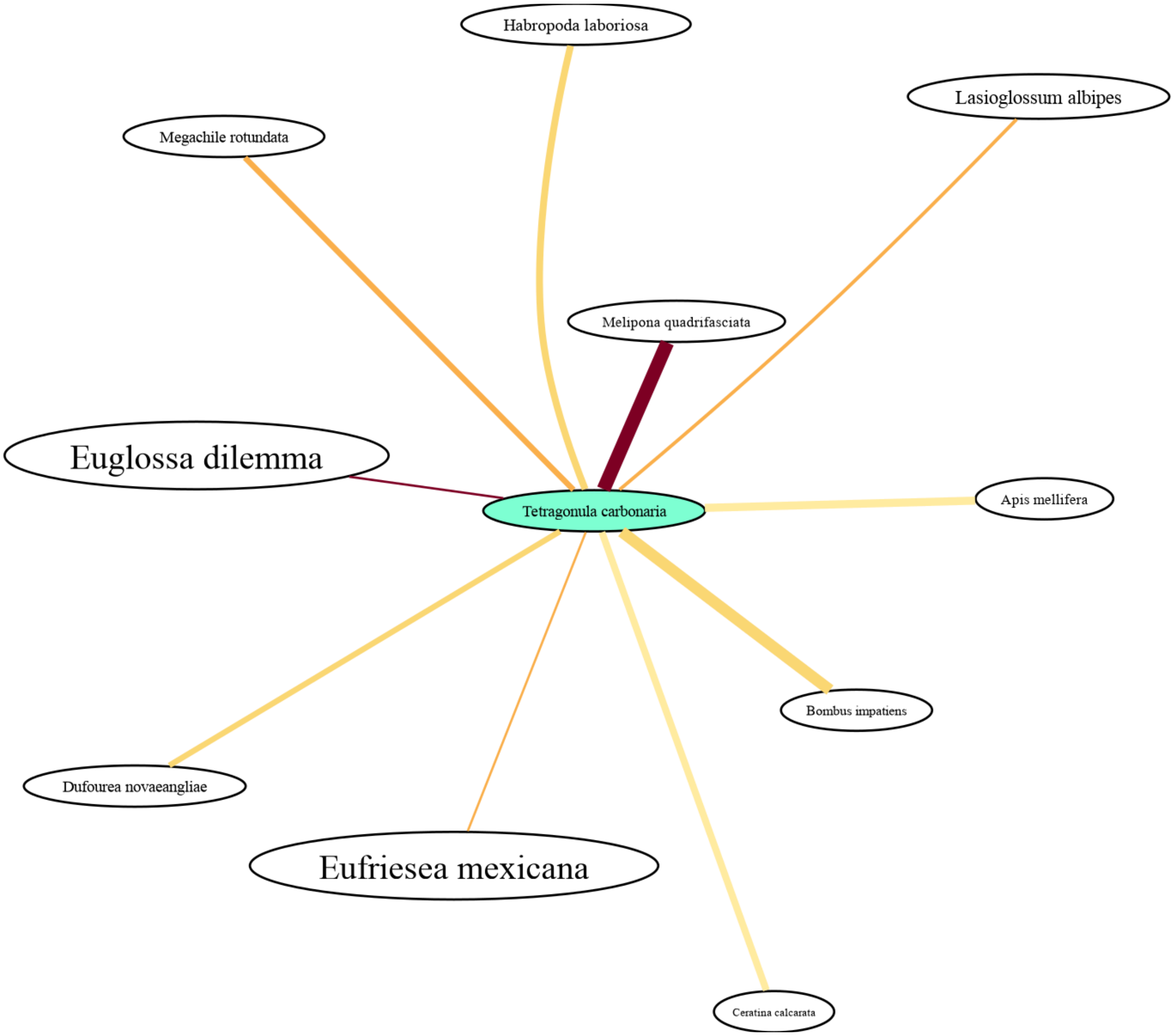


Figure 2.4: graph representation of the 11 bees Multispecies Whole Genome Alignment (MWGA). Each ellipse represents one assembly included in the alignment, with the reference assembly of *Tetragonula carbonaria* in the centre, highlighted in green. The size of each ellipse is correlated to the size of the corresponding assembly in bp. The physical distance between an ellipse and the central *Glossina morsitans* bubble is proportional to the phylogenetic distance between the corresponding species. The thickness of the connection between two ellipses represents the proportion of bases from the non-reference assemblies that were aligned to the reference assembly in the MWGA, and the colour of the connection indicates the total runtime of the pairwise alignment step between the two corresponding assemblies (yellow meaning faster, and red meaning slower).

Our role in this collaborative project was to provide evolutionary metrics for all genes in the new *Tetragonula carbonaria* assembly; these metrics were computed at several taxonomic levels by a member of our group who implemented a workflow dedicated to this task. We computed the 11 bees MWGA to estimate two metrics measuring conservation of genomic sequence between the aligned species for all the genes in the annotation for *Tetragonula carbonaria*, using the *mwgaw-analyses* workflow described in **Chapter 3**. Briefly, we computed *alignability* per nucleotide with *mwga-utils* and an evolutionary conservation score per nucleotide with phastCons, and these scores were averaged over the coding parts of each gene to obtain gene-level metrics. Together with other evolutionary metrics, sequence conservation can be used to make informed predictions on gene functional constraints (Ruzzante et al., 2022), the collaborators are exploring the use of these data in the context of sperm evolution in these stingless bees compared with other bees.

Discussion

At the start of this PhD project, all published Multispecies Whole Genome Alignments, either from studies of conservation of sequence or published as resources by the UCSC, were computed using the UCSC workflow, but this workflow was not publicly available or even described in a way which made it easily usable outside of the UCSC GenomeBrowser ecosystem. This lack of availability was a major reason for the lack of adoption of MWGAs as a resource in comparative genomics studies and genome annotation, which in turn inhibited methodological advances in MWGA computation and downstream analyses of sequence conservation, thus initiating a negative feedback loop. In light of this observation at the start of this project, we decided to first focus our efforts on implementing our own version of the UCSC workflow using modern workflow development tools and practices. To this aim, we leveraged the strengths of the workflow management system Snakemake as well as the package ecosystem Conda to develop *mwgaw-align*, described in this chapter. The result of this effort is a reproducible, documented, scalable, and modern implementation of the UCSC workflow to compute MWGAs, addressing the issue of quasi-unavailability of documented and reproducible approaches for this task.

The main outcome of the work presented in this chapter is the release of a usable workflow to facilitate the computation of new MWGAs. The usability of our workflow was validated by several users who were not familiar with alignments, and were able to compute new MWGAs with minimal effort for several clades, including stick insects (*Timema*) and fruit flies (*Drosophila*). The latter in particular included 36 assemblies and was used to compute coding-potential scores and identify putative novel functional elements using another workflow developed as part of this thesis work and detailed in **Chapter 3**. In addition to these practical outcomes, our work on *mwgaw-align* provides several benefits for the wider comparative genomics research community. First, because the vast majority of MWGAs computed before the start of this project resulted from the UCSC workflow, our implementation contributes, in an indirect way, to improving the reproducibility of these existing results. Indeed, it is now easier for users interested in these alignments to re-compute them, test the effect of parameter values for key steps in the workflow, and update the alignments with new assemblies if desired. Second, the modern and documented implementation of the workflow greatly facilitates testing improvements for some of the crucial steps of the workflow. Indeed, the Snakemake implementation enables users familiar with workflow management systems to easily remove and add steps in the workflow, or to replace a software used in a step with another software. For example, the pairwise alignment process is handled by LastZ, which was first released as BlastZ in 2000 (Schwartz et al., 2000); although LastZ has received considerable improvements since its release and is still widely used in genome alignments, new software to align genomic sequences has been released since then. One example of such software is minimap2 (Li, 2018), which aligns large genomic sequences efficiently. Our implementation of *mwgaw-align* effectively provides interfaces for crucial steps of the alignment process and would therefore enable testing the use of minimap2 over LastZ with minimal efforts, namely some additional format conversion scripts.

One limitation of the work presented in this chapter is that although we demonstrate the efficient applicability of the *mwgaw-align* workflow to several sets of genome assemblies, we did not complete a comprehensive testing of its scalability to hundreds of species. The 3-day runtime hard-limit on the UNIL compute facilities meant that such testing of scalability was not realistically achievable using the resources available to us. Given the investment in batching and parallelisation we believe that the pairwise alignment steps should be well-designed to scale to hundreds of species. However, from the results presented for our computed MWGAs it is clear that peak memory usage occurs during the computation of the MWGA from the pairwise alignments. Although we have not been restricted by this memory usage for the datasets computed in this project, we anticipated two solutions to this potential problem when scaling to larger datasets. First, the memory-intensive step can be specifically run on high-memory nodes, which can provide five to ten times more memory than the peak recorded in the computation of our datasets. Second, the computation of the MWGA from pairwise alignments can be parallelised by splitting the reference assemblies into regions - for instance, using the batches already computed in the workflow - and computing the MWGA for each region separately before merging all the resulting MWGAs into a final, global MWGA. The alternative solution of redesigning the MULTIZ merging algorithm was considered beyond the scope of this thesis.

Although the UCSC workflow remained the standard approach to compute MWGAs when we started implementing *mwgaw-align*, the UCSC GenomeBrowser team had already published Cactus as an alternative workflow to align genomes. Cactus was first published in 2011 in a methodological paper showcasing the power of a new data structure, *cactus graphs*, to represent genome alignments (Paten et al., 2011); in this first paper, Cactus is shown to have similar precision to MULTIZ but a higher recall on test datasets, making it a promising alternative to these approaches. However, as usual at the time, Cactus was released only as unpackaged and uncompiled software with limited documentation (https://github.com/ComparativeGenomicsToolkit/cactus/tree/genome_research_alignment_paper), and was thus difficult to adopt for users outside its development team. Consequently, there were no MWGAs computed using Cactus following this first publication, and in fact, the software saw little improvements for a long time. It is only years later that a new team - still within UCSC - took over the development of Cactus, with the first modern release published in 2020 (Armstrong et al., 2020). The software is now under active development and provides notable improvements over the classic UCSC workflow: 1) it implements a reference-free format for MWGAs, HAL (**H**ierarchical **AL**ignment format, Hickey et al., 2013), 2) it incorporates many intermediary steps to refine pairwise alignments, thus generally producing better alignment results than the classical workflow, and 3) it is released as a docker image, making it actually available to external users. Thanks to these improvements and to the active development efforts still ongoing to improve both the alignment process and usability, Cactus established itself as the default workflow to compute MWGAs and will most likely solidify this status in the coming years, as illustrated in a recent publication computing large-scale alignments (Christmas et al., 2023; Zoonomia Consortium, 2020).

However, despite these improvements and the recent success of the Cactus development team in computing large scale MWGAs, the software is not without some drawbacks. First, despite a much improved availability since its earlier releases, Cactus is still lacking in user-friendliness and usability. Second, our workflow enables the production of MAF files that are the common starting point for many downstream analyses, where the desired results are meant to be constrained to the coordinate system of a single reference species assembly (several of these are described in **Chapter 3**). Although the reference-free HAL format implemented in Cactus is intuitively a great improvement, almost all downstream alignment processing and analysis tools still require a referenced-based alignment, so the HAL data still need to be transformed into the traditional MAF file; the Cactus development team has started implementing reference-free versions of some of these downstream processing tools, but most of them remain only available for reference-based alignments. Third, although it should be possible to update an assembly in a Cactus-generated alignment without re-running the entire pipeline, this process is currently complicated for users to do and requires considerable manual intervention. Our *mwgaw-align* workflow was designed with user-friendliness in mind, making it relatively simple to set up and run, especially for experienced Snakemake users but also for relative novices. The workflow management system also greatly simplifies the task of updating existing MWGAs, allowing for the addition, removal, or replacement of assemblies with subsequent execution of only the required steps to recompute the new MWGA (*i.e.* avoiding re-running many already computed pairwise alignments). Fourth, as we illustrated in this chapter, the optimisations we brought to *mwgaw-align* enable efficient computation of MWGAs including more than 20 assemblies. In contrast, the extra steps implemented in Cactus to improve the alignment process greatly increase the computational cost of generating MWGAs; in fact, in a comparison with an alternative alignment tool for closely-related assemblies, Sibelia (Minkin & Medvedev, 2020), the authors noted that Cactus was not able to align eight mouse genomes in seven days. Finally, the implementation of *mwgaw-align* will contribute to the ongoing technical development of Cactus itself, as our workflow enables the streamlined generation of MULTIZ-based MWGAs that can be used to benchmark the outputs from the Cactus approach.

In the field of computational biology it is often observed that healthy competition amongst methods designed to solve similar challenges promotes advances in accuracy and performance. Solving the technical challenge of multiple whole genome alignments is no exception, as evidenced by the 12 participant methods in the Alignathon initiative (Earl et al., 2014). Until the development of Cactus, the MULTIZ-based approach developed by the UCSC team was arguably the only practically feasible approach with good accuracy and performance, albeit only usable essentially by the UCSC team themselves. This thesis work to build a reproducible, portable, and scalable workflow for running the MULTIZ-based approach is therefore an important contribution to the field as a whole. Instead of being consigned to history as a near-unusable and poorly documented set of utilities and complex wrappers, the MULTIZ-based approach was given a new lease of life by our *mwgaw-align* workflow, enabling its continued use for generating high-quality MWGAs, and possibly for participation in future Alignathon initiatives.

Chapter 3: Computing metrics from a Multispecies Whole Genome Alignment with *mwgaw-analyses*

Summary

This chapter describes the *mwgaw-analyses* workflow that facilitates the downstream analysis of multispecies whole genome alignments (MWGAs), interfacing with the *mwgaw-align* workflow described in **Chapter 2** to build MWGAs. We first describe the types of analyses implemented in the workflow for computing estimates of sequence conservation and protein-coding potential. We then detail the methodological implementations for input data processing, computing conservation and coding-potential metrics, creating gene-level summaries, and building data track hubs for visualisations using genome browser platforms. The development of supporting tools and utilities are detailed in **Chapter 5**. The results section then summarises the types of outputs produced by the workflow, highlighting the data visualisation options from nucleotide-level to gene-level and even genome-wide data summaries. Importantly, the *mwgaw-analyses* workflow's primary input is a MWGA in MAF format, meaning that it can be applied to MWGAs produced by other workflows, e.g. Cactus, and it is not dependent on our *mwgaw-align* workflow to build the required MWGA.

Introduction

MWGAs are a powerful resource for any analysis comparing multiple genomes. In the context of this thesis, we want to use MWGAs to 1) compare patterns of sequence conservation for different families of functional genomic elements across arthropods, and 2) generate clues to assist the annotation of coding and noncoding functional elements in assemblies. Both these goals require computing metrics to estimate conservation of sequences across the alignment to identify conserved regions in the reference assembly.

A first and naive approach to explore sequence conservation in a MWGA is to compute simple metrics like the proportion of aligned assemblies (**alignability**) and the proportion of assemblies with the same nucleotide as the reference (**identity**) for all positions in the reference assembly. These metrics do not consider the evolutionary history of species in the alignment and thus cannot be used to accurately estimate conservation or constraint of sequences among these species. However, the information they provide can help understand some of the patterns potentially observed in more complex estimates of sequence conservation. We implemented the computation of alignability and identity in the *metrics* utility of our *mwga-utils* software suite (technical description presented in **Chapter 5**).

Accurately estimating sequence conservation requires reconstructing the evolutionary relationship between species included in the alignment. In practice, this task involves the use of a statistical model taking into account substitution rates in each aligned species while accounting for patterns in the mutational process, *e.g.* transitions being more frequent than transversions. When genome sequencing and computational advances made MWGAs possible in the early 2000s, several tools were developed to identify conserved elements in MWGAs e.g. (Margulies et al., 2003; Ovcharenko et al., 2005), yet few of these methods actually used an evolutionary model, but see (Boffelli et al., 2003). The first software to properly implement a statistical model taking into account phylogenetic relationships between aligned species and substitution patterns was phastCons (Siepel et al., 2005), which was quickly adopted as the standard to identify conserved elements and has been used in most studies investigating conservation of sequence since then (Hecker & Hiller, 2020; Hupalo & Kern, 2013; Lindblad-Toh et al., 2011; Stark, Lin, et al., 2007) and in conservation tracks from the UCSC Genome Browser (Miller et al., 2007). PhastCons implements a phylogenetic Hidden Markov Model (phylo-HMM) with a state for conserved regions and a state for non-conserved regions. Both states are described by the same phylogenetic model, with the average substitution rate in conserved regions expressed as a fraction of the average substitution rate in non-conserved regions. A defining feature of phastCons is that all model parameters (*e.g.* branch lengths, substitution rates, nucleotide matrix, …) are inferred from the data by maximum likelihood except for two constraints: expected proportion of conserved bases in the reference assembly, and average length of conserved elements. Models estimated from the MWGA are then used to identify conserved elements using a variable-size sliding window and to compute a conservation score for each base in the reference assembly; in practice, these conservation scores represent the probability that a site was generated by the conserved state of the model and thus varies between 0 and 1. Phylo-HMMs enable phastCons to integrate the phylogenetic relationships among aligned species and complex substitution models allowing multiple substitutions, and to efficiently identify conserved elements.

Identifying conserved elements is the first major step to generate annotation clues for the reference assembly. This information can be supplemented with additional metrics to infer the functional role of identified conserved elements; a crucial distinction between types of functional elements is whether these elements encode proteins. Most assemblies undergo an annotation process to identify protein-coding functional elements, *i.e.* protein-coding genes. Often, RNA-seq data can be used to delimitate the boundaries of these genes and data from functional studies can be added to curate the annotation. However, in many cases, protein-coding genes are identified *de novo* using only homology to genes in other species and patterns of nucleotide sequence. Annotations generated by this automated process are a valuable resource, but they are often incomplete and some genes may not be properly delimited. Specific patterns of conservation of sequence in MWGAs can provide information to 1) determine whether a gene encodes a protein, 2) delimit the boundaries of annotated genes, and 3) identify new genes missed by automated annotation. A particular method looking at Codon Substitution Frequencies (CSF) has been implemented in the software PhyloCSF (Lin et al., 2011), which has been successfully used to identify and characterise protein-coding genes in a diversity of species (Khan et al., 2020; Lian et al., 2018; Mudge et al., 2019). The approach implemented by PhyloCSF is very similar to that of phastCons: PhyloCSF also uses two phylogenetic models, one for coding regions, and one for non-coding regions. For each codon, PhyloCSF then estimates the probability that the sequence is generated by the coding model based on the log-likelihood ratio of the two models. In contrast to phastCons, however, PhyloCSF relies on complex empirical codon models (ECMs) inferred from alignments of coding and non-coding regions using thousands of parameters. To be accurate, these ECMs require precisely annotated coding regions, and therefore PhyloCSF provides curated ECMs for several datasets; based on a personal communication with the author of PhyloCSF, these curated models are much favoured over models generated using new data without curation, and thus the most efficient way to use PhyloCSF is to adjust curated models to the new input data.

In practice, all these analyses involve many steps and greatly benefit from parallelisation to optimise user runtime. We efficiently implemented these steps in our automated but customisable *mwgaw-analyses* workflow, which computes multiple metrics and outputs from a multispecies whole genome alignment. Base metrics - alignability and identity - are estimated only at the nucleotide level, but both phastCons and PhyloCSF can generate two types of outputs: a raw score per nucleotide (for phastCons) or per codon (for PhyloCSF), and a list of most conserved (for phastCons) or putative coding (for PhyloCSF) elements. Nucleotide-level data for base metrics and phastCons scores can be averaged over the length of annotated coding regions for each gene in the reference assembly to generate gene-level metrics, which can be useful to understand the gene's evolutionary trajectory and potentially inform on its function (Ruzzante et al., 2022); in **Chapter 4**, we use our gene-level metrics to characterise families of immune-related genes in mosquitoes. Finally, the amount of data generated by these analyses can be daunting; to help understand and make sense of these results, we focused on implementing several visualisation approaches at multiple levels. First, a GenomeBrowser track is generated for each output - at the nucleotide, element, and gene level, in order to visualise the associated metric value along the assembly for the reference species; this visualisation is a powerful tool to explore patterns of a chosen metric along a region of interest, or to visually identify new regions of interest. These tracks are integrated in a GenomeBrowser track hub, which provides a convenient way to load multiple tracks and associated metadata in a GenomeBrowser instance. Second, gene-level metrics are integrated in a web interface displaying 1) all gene-level metrics in a table with information on the gene and potential links to an alignment viewer (CodAlignView), and 2) figures showing sequence conservation metrics per nucleotide along with optional information (*e.g.* gene model, or sequences marked as repeated) across the length of each gene in the reference annotation.

The automated and customisable *mwgaw-analyses* workflow described in this chapter provides the tools required to turn the "raw data" contained in a MWGA into quantified metrics that describe nucleotide level sequence conservation along the length of the genome, which can also be averaged over any annotated features such as exons or genes. Importantly, the results are organised and formatted so they can be easily integrated into widely-used genome data visualisation platforms, enabling users to explore the results as data tracks in genome browsers.

Methods

Processing the MWGA, annotation, and phylogeny inputs

The *mwgaw-analyses* workflow uses as the main input an MWGA in standard MAF format, for instance one generated using our *mwgaw-align* workflow or one transformed from the HAL format produced by Cactus. In addition, the workflow requires an assembly file and an annotation file for the reference assembly as well as a phylogenetic tree defining the relationships between species included in the alignment, including branch lengths which are required to compute coding potential with PhyloCSF. The reference assembly fasta file is processed as it would be for inclusion in the MWGA generation workflow, i.e. it is formatted, converted to a binary format, indexed, and partitioned into batches as detailed in the Methods section of **Chapter 2**. This partitioning is required to parallelise the costly computation of per-base sequence conservation and coding potential scores by splitting the MWGA into batches which can be processed concurrently. The annotation file, in General Feature Format (GFF), for the reference assembly is re-formatted following the same rules as the fasta file for the reference assembly, and is split into scaffold-level annotation files. These are required to be able to average per-nucleotide metrics over any selected annotated features, such as exons or whole genes. Finally, while PhyloCSF uses the phylogeny as provided, the tree provided is converted to a topology-only dendrogram format required by phastCons, *i.e.* (species1,species2,(species3,species4)).

Computing per-base alignment metrics

In addition to conservation score (phastCons) and coding potential (PhyloCSF), basic metrics are generated for each nucleotide in the reference assembly using our *metrics* software from *mwga-utils* (technical description presented in **Chapter 5**). This software computes two metrics: **alignability**, defined as the proportion of assemblies that were aligned at a position in the reference assembly, and **identity**, defined as the proportion of assemblies with the same nucleotide as the reference at a position. These basic metrics can help understand and interpret some of the patterns observed for conservation score and coding potential. The design and implementation of *mwga-utils* are explained in **Chapter 5**; briefly, the software parses the MAF file to extract nucleotide information in all assemblies for each genomic position in the reference assembly. This information is used to compute alignability and identity for the current position, and these scores are stored to be exported after processing the entire input MAF file. The output is a WIG file (https://genome.ucsc.edu/goldenPath/help/wiggle.html) for each computed metric, containing the computed value for each position in the reference assembly, including a *0* for positions where metrics could not be computed, *e.g.* regions where no assembly was aligned to the reference.

Estimating sequence conservation with phastCons

Conservation of sequence per nucleotide is estimated with phastCons, which relies on two phylogenetic models to compute conservation scores, one for conserved regions and one for non-conserved regions. The recommended way to estimate these models for alignments including more than a few species is to 1) infer the phylogenetic model for non-conserved regions from fourfold degenerate (4d) sites in coding regions, *i.e.* sites forming the third base of a codon for which any base will encode the same amino acid, and 2) infer the phylogenetic model for conserved regions from conserved sites, usually the first position in codons. This method requires information on coding regions in the reference assembly, which is inferred from an annotation file supplied by the user in the workflow's config file. The first step of this process is to extract all 4d sites and 1st codon positions from the alignments. First, the corresponding genomic region is extracted from the input MWGA for each alignment batch using the *maf_parse* software from the PHAST suite, and features for the corresponding region are extracted from the user-supplied annotation file using a Python script. The *msa_view* software from PHAST is then used to identify 4d sites and first codon positions in the alignments; details of this procedure are implemented in the wrapper script *phastCons_models_batch.py* which processes all sequences in a batch. First codon positions and 4d sites for all batches are respectively merged into a single file each, and used to estimate the phylogenetic model for conserved and non-conserved sites with the *phyloFit* software from PHAST. Model estimation requires a tree describing the relationships among species, which is generated from the user-supplied Newick tree using a Python script. Finally, conservation scores are computed for each batch with phastCons, and results from all batches are concatenated to create a WIG file of conservation scores for all bases in the reference assembly.

Per-nucleotide conservation scores are used to compute a list of most conserved elements using the *--most-conserved* mode in phastCons. This mode uses the WIG files generated in the previous step to identify continuous regions of high sequence conservation, based on two user-supplied parameter values: estimated length of a conserved element, and estimated coverage of conserved elements in the reference assembly. The process is parallelised by running phastCons on each contig separately, and the output is a Browser Extensible Data (BED) format file of conserved elements for each contig in the reference assembly. These BED files are then aggregated into a single file containing all conserved elements identified in the alignment.

Estimating coding potential with PhyloCSF

Coding potential per nucleotide is estimated with PhyloCSF (Lin et al., 2011); similar to phastCons, this software relies on models to estimate scores. PhyloCSF models consist of three files: 1) a phylogenetic tree in Newick format with branch lengths describing the relationship between species included in the model, 2) a table describing transition probabilities between all possible codon pairs in the coding-sequence model model, and 3) a table describing transition probabilities between all possible codon pairs in the non-coding-sequence model. Unlike phastCons models, however, models used by PhyloCSF require fine tuning to perform accurately, and we learned from communicating with the main author of PhyloCSF that using the model developed for a dataset of 12 *Drosophila* species (*12flies*) would yield better results than estimating a new model with PhyloCSF for subsets of arthropod species. In practice, this would require creating a new model using the same transition probability as the *12flies* model, but replacing the phylogenetic tree with one describing the relationship between species in the focal subset of species, *i.e.* species included in the input alignment. Consequently, the first step in running PhyloCSF in our analyses workflow is creating a dummy model in the *datasets* folder of PhyloCSF by copying the user-specified Newick tree and the coding and non-coding models from the *12flies* dataset into a new model.

PhyloCSF is then run using this model to estimate per-base coding potential for each of the three forward and three reverse reading frames for all positions in the reference assembly. To do so, each alignment batch is first extracted from the MWGA with *maf_parse* and is then split into overlapping 3bp alignments representing potential codons. Because this process can generate millions of temporary files, thus overloading the file system, each batch is processed in successive sub-batches of 1,000 bp to limit the number of concurrent temporary codon files. PhyloCSF is run on all codon files for each sub-batch using a fixed strategy (*--strategy=fixed*) to optimise compute time and non informative output columns are filtered out on the fly to further optimise disk usage. Output files for all batches are merged and processed with a Python script to generate six final WIG files, one for each of the six reading frames.

Coding potential scores per codon can be used to generate a list of putative coding elements, similar to phastCons' list of most conserved elements. This approach was used to identify 353 and 51 cases of functional readthrough of stop codons in *Anopheles gambiae* and *Drosophila melanogaster* (Jungreis et al., 2016), respectively, and to add 144 conserved protein-coding genes to the human GENCODE gene set along with additional coding regions within 236 previously annotated protein-coding genes (Mudge et al., 2019). However, the actual method to produce lists of high-scoring putative coding elements was never directly implemented in PhyloCSF. Instead, these results were achieved using a collection of unreleased scripts that were implemented and designed for the author's working environment. After a personal communication, the author agreed to share these scripts and include them in our workflow. The vast majority of the work to convert these scripts from Python 2.7 to Python 3, update the dependencies, and integrate them in the framework of *mwgaw-analyses* was performed by another member of the lab and will not be described here. I personally assisted with integrating this updated utility into the workflow, and *mwgaw-analyses* can thus produce a list of putative coding elements as an output.

Computing per-gene metrics for the reference assembly

The workflow computes alignability, identity, and conservation scores for each genomic position in the reference assembly, including coding and non-coding regions. To this day, many biological questions still focus on coding regions, since they are the regions best associated with a known biological function and most easily comparable between individuals and species. When an annotation detailing coding regions is available for the reference assembly, usually a GFF file, the workflow can compute the average value of each of three aforementioned metrics over the length of each gene. The resulting per-gene metrics can be used to inform evolutionary characterisations of genes and gene families, i.e. quantifying evolutionary sequence conservation per gene to be able to distinguish between functional categories of genes that are normally highly conserved/constrained and functional categories of genes that show high variability in their sequences over evolutionary time. An example of the application of gene-level metrics to evolutionary questions is presented in **Chapter 4**.

The first step in computing gene-level metrics is loading the information contained in the GFF annotation file into a custom GFFFile Python object. This object contains all important information from the GFF for each annotated gene, as well as an index associating each genomic position in the reference assembly with the relevant gene. Then, for each metric, the WIG file of per-base metric values generated in one of the previous steps is read position by position. When the current genomic position falls within a gene, as indicated by the index created while loading the annotation file, gene-level metric values are incremented for the relevant gene in a data structure. After processing the entire file, incremented metric values are divided by the length of the gene for each gene in the annotation. These averaged values are exported to a tabulated file containing the gene name and average value of each computed metric on each line.

To facilitate exploring the computed metrics and identifying genes of interest, the resulting tabulated file can be converted into an interactive web table that can be filtered, ordered, and searched. This table also links each gene to the corresponding genomic region in a CodAlignView instance (a web-based tool developed by the PhyloCSF developers), which visualises the alignment used to compute the metrics. To generate this table, we implemented a Python workflow which takes as input the tabulated file of average metric values for each gene and automatically generates a static website with the interactive table. This workflow is available at https://gitlab.com/evogenlab/mwga/mwga-website and allows to easily browse gene metrics results for any alignment computed with our *mwgaw-analyses*.

Creating a Genome Browser track hub

The output generated by the analyses performed in the workflow consists of WIG files and BED files; these file formats are standard for use in downstream analyses tools and easily parsable, but they are not easy to browse to visualise results. To assist with visualisation and exploration of analyses results, a Genome Browser track hub is generated for all metrics specified by the user in the config file. In practice, a track hub is a folder that contains the reference assembly in 2bit format, an index file for this assembly, a data track file for each metric, and two files describing the tracks. The reference assembly and index files are copied from the first steps of the workflow; a data track file in bigWig is obtained from the final WIG output for each metric with the wigToBigWig utility from UCSC kent utilities, or from the final BED output using bedToBigBed from the same software package. The two files describing the track hub are generated using our py-GB-tracks python package (https://pypi.org/project/py-GB-tracks/) from a track description YAML file provided with the workflow; tracks can be customised by editing this YAML file. Implementation details for the py-GB-tracks package are provided in **Chapter 5**. The resulting track hub can be automatically loaded in a genome browser (GenomeBrowser or JBrowse) instance to visualise conservation scores, coding potential scores, and putative conserved / coding elements.

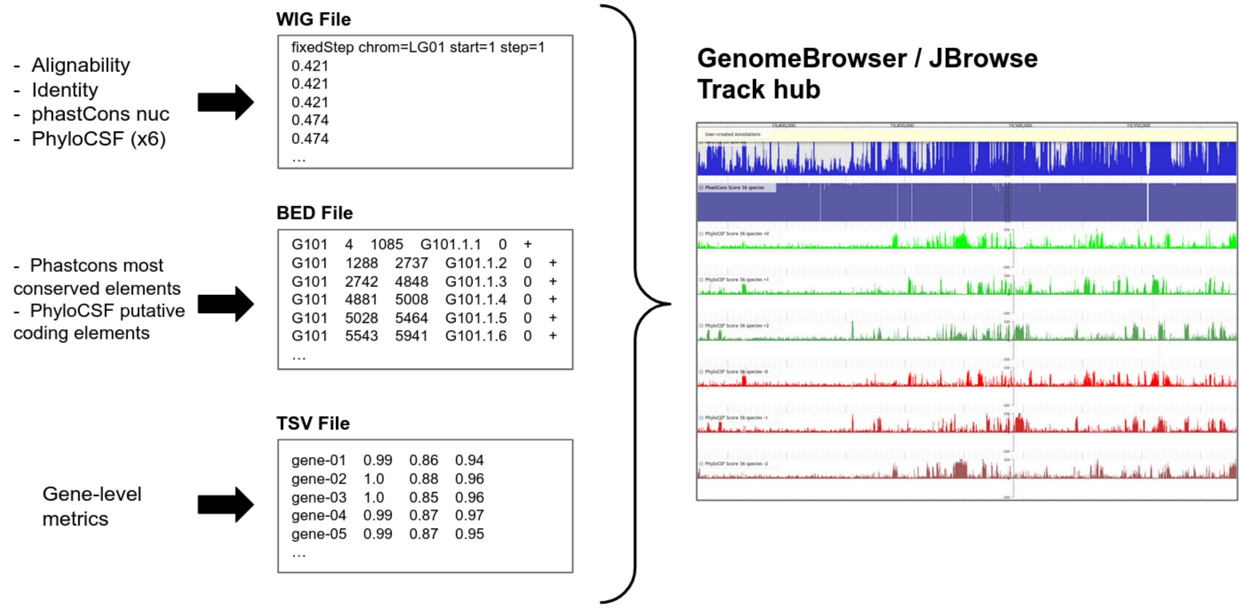
Reproducible, portable, and scalable implementation of the workflow

Similar to the *mwgaw-align* workflow, the *mwgaw-analyses* workflow is implemented using the Snakemake workflow management engine (Köster & Rahmann, 2012). Conda environments were implemented to handle dependencies for all steps in the workflow; all pieces of software used to perform analyses were either already available in the Bioconda channel (Grüning et al., 2018) or were packaged by us if missing (a detailed list of packaged software is provided in **Chapter 5**), making the workflow entirely portable on all platforms supported by Conda. The workflow was implemented following Snakemake's official guidelines for workflow organisation and implementation, making it easy to use by experienced Snakemake users and efficient to review by experienced Snakemake developers. The config file, repository, and implementation are thoroughly documented to promote transparency and encourage improvements by the community. Performance-critical steps in computing conservation scores and coding potential were parallelised, and a batching system was implemented to avoid overloading the file system with too many small files. These performance optimisations can reduce the runtime of the workflow by multiple days and are a key factor allowing to scale the workflow to large MWGAs. We implemented detailed resources requirements for each step of the workflow (see **Chapter 5** for a description of the resources system); coupled with the portability of Conda environment, this enables executing the workflow on a local computer, a computational platform, or a cloud-based environment with minimal efforts.

Results

Overview of the *mwgaw-analyses* workflow output

The final output of running the *mwgaw-analyses* workflow is a series of files for each of the computed metrics, and a track hub output that organises these files into an easy to load package for visualisations of the tracks in GenomeBrowser or JBrowse, these outputs are summarised in **Figure 3.1**. The first output generated by each analysis is a "wiggle" (WIG) file which contains the value of the corresponding metric along the genomic coordinates of the reference assembly. These per-nucleotide metrics are provided for each genomic position for Alignability, Identity, and phastCons, and for each codon (whether it is a real codon or not) in each of the six reading frames for PhyloCSF. The WIG files are used by phastCons and PhyloCSF to generate a set of most conserved elements and putative coding elements, respectively, which are exported as Browser Extensible Data (BED) format files - one element per line with the corresponding genomic location and associated score. The nine WIG files and two BED files are integrated in the track hub, along with metadata to assist their visualisation when loaded in a genome browser. Finally, the WIG files for Alignability, Identity, and phastCons are used to compute per-gene average metric values, which are exported as a tabulated file as well as a browsable HTML table to explore the data.


- Alignability
- Identity
- phastCons nuc
- PhyloCSF (x6)
WIG File
- Phastcons most conserved elements
- PhyloCSF putative coding elements
BED File
Gene-level metrics
TSV File
GenomeBrowser / JBrowse Track hub

```
fixedStep chrom=LG01 start=1 step=1
0.421
0.421
0.421
0.474
0.474
...
```

```
G101   4    1085   G101.1.1   0   +
G101   1288   2737   G101.1.2   0   +
G101   2742   4848   G101.1.3   0   +
G101   4881   5008   G101.1.4   0   +
G101   5028   5464   G101.1.5   0   +
G101   5543   5941   G101.1.6   0   +
...
```

```
gene-01   0.99   0.86   0.94
gene-02   1.0    0.88   0.96
gene-03   1.0    0.85   0.96
gene-04   0.99   0.87   0.97
gene-05   0.99   0.87   0.95
...
```

Figure 3.1: summary of the output of *mwgaw-analyses*. Per-nucleotide values for alignability, identity, phastCons conservation score, and PhyloCSF coding potential are produced as WIG format files, while the most conserved elements and the putative coding elements are output as BED format files. Finally, gene-level metrics are first output as a TSV file and converted to BED format files for future display. A GenomeBrowser track is created for each of these metrics and the tracks are packaged together as a track hub for convenient loading into GenomeBrowser or JBrowse. The visualisation shows an example genomic region displaying tracks from top to bottom: Alignability scores, phastCons scores, PhyloCSF scores for the six reading frames (forward in green and reverse in red).

Exploring and visualising the output

GenomeBrowser track hub

The majority of the output of *mwgaw-analyses* - and the output that has to be computed first - consists of per-nucleotide (or per-codon) metric values along the entire length of the reference assemblies, exported as WIG files. This output is low-level and comprehensive, but because of its size and format, it is not easily browsed and visualised by users. Yet, when processing such a large amount of complex data, visualisation is a powerful tool to explore global patterns and to better understand local patterns identified with quantitative analyses. A major medium to visualise data along a genome takes the form of a genome browser, a usually online-based platform allowing to navigate across a genome sequence while displaying multiple information tracks - metrics, annotations, etc. One of the first genome browsers was released to support exploration of the first assembled human genome, highlighting the importance of such a tool in understanding large amounts of data. Modern genome browsers implement loading multiple related tracks together through the use of track hubs, and one of the last steps of *mwgaw-analyses* generates a track hub for all the output computed by the workflow. The track hub generated for the MWGA of 36 *Drosophila* assemblies after loading in the modern genome browser JBrowse is displayed in **Figure 3.2.A**. The panel shows phastCons scores, the six-frame PhyloCSF coding potential scores, and genome annotations for a region spanning from 19.35 Mbp to 19.6 Mbp on scaffold NC_04628.1 in the assembly of *Drosophila albomicans*. This region harbours three genes, and therefore conservation of sequence is generally high across the entire window, albeit lower in untranslated regions and heterogeneous in some parts of the exons (blue track). Exons in all three genes are matched with regions of high coding potential, each located on different reading frames; it is worth noting that a region of high coding potential on one reading frame corresponding to the actual reading frame on which the exon is translated is mirrored by a region of high - but lower - coding potential on the matching reverse reading frame, because of conserved codon sequences.

The track hub generated for the MWGA of 22 mosquito assemblies after loading in VectorBase's GenomeBrowser instance is displayed in **Figure 3.2.B**, which shows phastCons scores per nucleotide (in yellow) as well as PhyloCSF coding potential scores for two forward reading frame (red) and one reverse reading frame (blue) and genome annotations for a region between 41.257 and 41.260 Mbp on chromosome 2L in the assembly of *Anopheles gambiae*, which contains the gene AGAP007033. The tracks reveal a relatively low level of sequence conservation and coding potential for this gene, which can happen for fast-evolving genes (see **Chapter 4**). In contrast, a region a few Kbp upstream of AGAP007033, which does not match any gene in the annotation, exhibits high overall sequence conservation as well as several blocks of high coding potential. Available expression data for *Anopheles gambiae* revealed low but existing level of expression in this region; by comparing this expression data with the phastCons and PhyloCSF results, we were able to delimitate the structure of a gene in this region, which was submitted as a new annotation for this assembly. It is possible that automated annotation pipelines discarded this region despite existing expression data because of its close proximity to AGAP007033, which is highly expressed. This example illustrates how the exploration of MWGA metrics in specific genomic region, enabled by the track hub, can benefit genome annotation efforts.

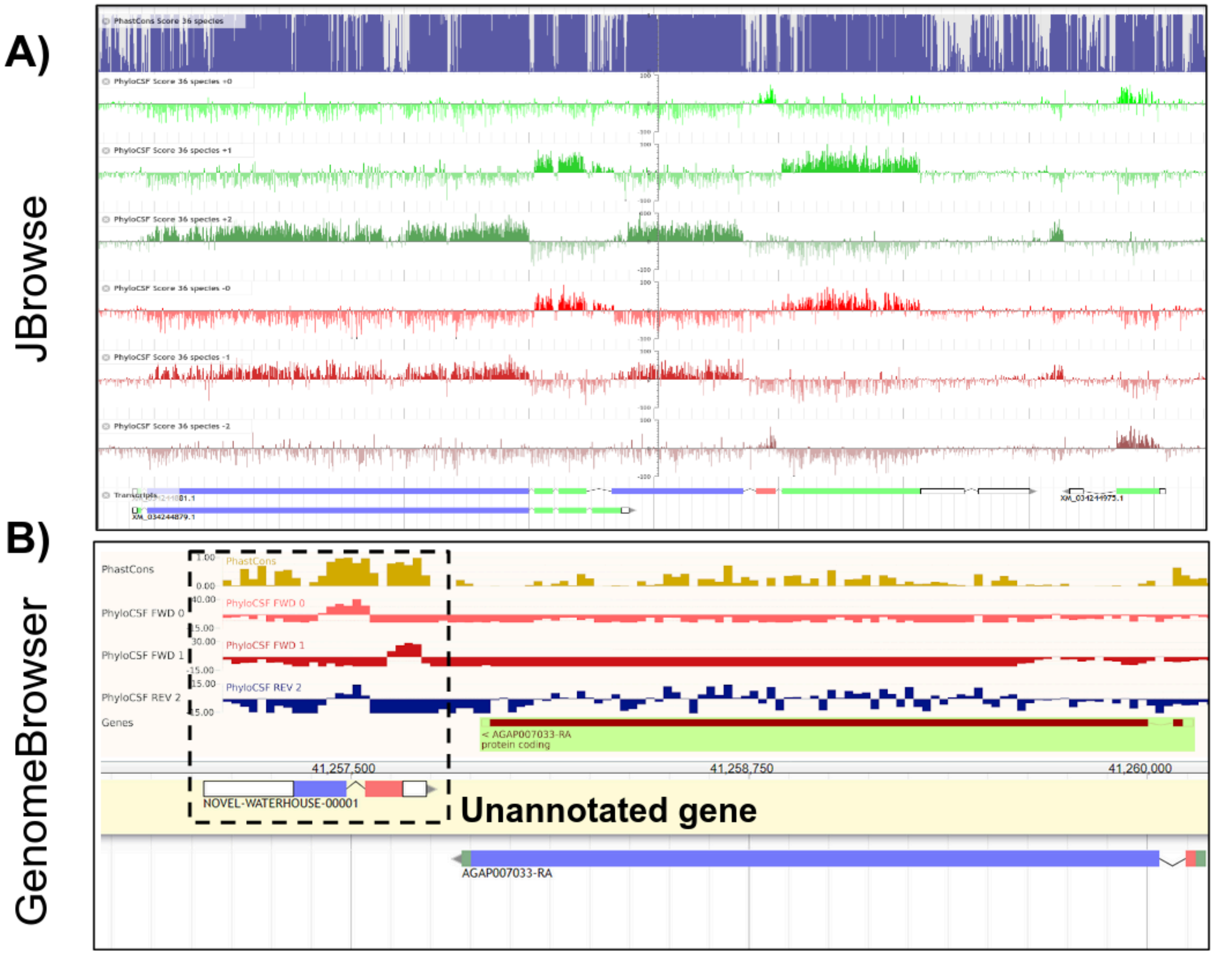
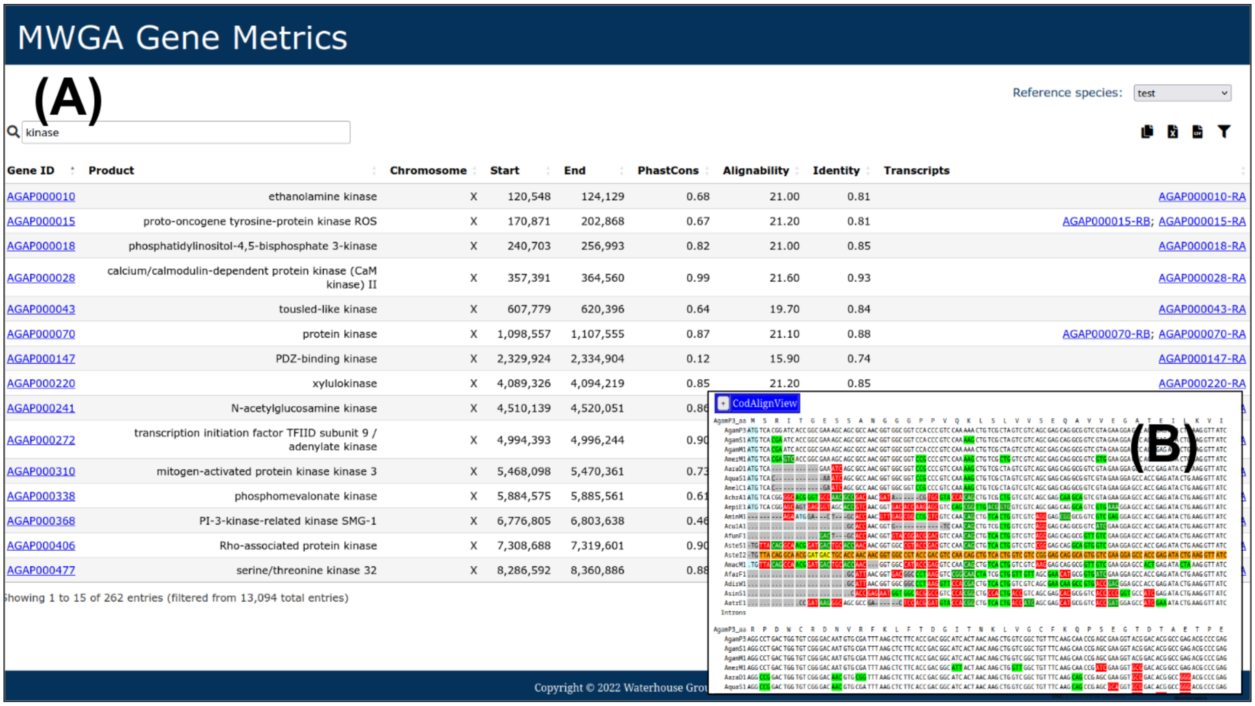


Figure 3.2: track hub visualisation on two different platforms. **A)** Track hub generated for a MWGA of 36 *Drosophila* genomes visualised in a JBrowse-based genome browser and annotator (Apollo), showing a region located between 19.35 Mbp and 19.6 Mbp on NC_04628.1 in the assembly of *Drosophila albomicans*. From top to bottom, the figure displays the following tracks: phastCons score per codon (blue), PhyloCSF coding potential per codon for forward reading frames (three green tracks), PhyloCSF coding potential per codon for reverse reading frames (three red tracks), and gene models from the annotation. The figure shows high coding potential on PhyloCSF tracks matching each exon in the annotated genes. **B)** Track hub generated for a MWGA of 22 mosquito species visualised in VectorBase's GenomeBrowser instance; the figure shows a region in the vicinity of the gene AGAP007033, between 41.257 and 41.260 Mbp on chromosome 2L in the assembly of *Anopheles gambiae*. From top to bottom, the following tracks are displayed: phastCons score per nucleotide (yellow), PhyloCSF coding potential per codon for two forward reading frames (in red), and PhyloCSF coding potential per codon for one reverse reading frame (in blue). The figure highlights a region showing peaks of phastCons score per nucleotide matched by two regions of high coding potential on the forward reading frames. After comparing this information with available RNA-seq data, it was determined that this region was likely an unannotated gene with two exons, which could have been missed because of its proximity to the highly expressed AGAP007033 gene.

Browsable table of gene-level metrics

MWGAs enable the exploration of sequence conservation and evolutionary constraints along the entire genome, including all types of genomic elements. Although an increasing number of studies attempt to unravel the role of non-coding functional elements in controlling biological functions, most of the focus in research remains on protein-coding genes, whose link to biological function is more straightforward and better understood; our work presented in **Chapter 4** highlights the power of gene-level sequence conservation metrics in understanding gene function. To assist with the exploration of gene-level metrics computed by *mwgaw-analyses*, we implemented an automated workflow to programmatically generate a browsable table of metrics, as illustrated in **Figure 3.3** for the 22 mosquito MWGA. This table can be filtered, for instance to focus on genes encoding a specific class of product - *kinases* in the figure; advanced filters are available to easily select genes of interest. In the example displayed in the figure, each gene ID is linked to the corresponding gene information page on VectorBase. In addition, the raw MWGA data for the genomic region around each transcript for each gene can be visualised using a CodAlignView instance (**Figure 3.3.B**), accessed by a hyperlink on the transcript ID. The browsable table paired with the CodAlignView visualisation is a powerful tool to extract sets of genes fulfilling conditions of interest to a user, and to explore patterns of evolutionary sequence conservation at the nucleotide level for a set of genes of interest.



MWGA Gene Metrics

kinase

Reference species: test

Gene ID	Product	Chromosome	Start	End	PhastCons	Alignability	Identity	Transcripts
AGAP000010	ethanolamine kinase	X	120,548	124,129	0.68	21.00	0.81	AGAP000010-RA
AGAP000015	proto-oncogene tyrosine-protein kinase ROS	X	170,871	202,868	0.67	21.20	0.81	AGAP000015-RB; AGAP000015-RA
AGAP000018	phosphatidylinositol-4,5-bisphosphate 3-kinase	X	240,703	256,993	0.82	21.00	0.85	AGAP000018-RA
AGAP000028	calcium/calmodulin-dependent protein kinase (CaM kinase) II	X	357,391	364,560	0.99	21.60	0.93	AGAP000028-RA
AGAP000043	tousled-like kinase	X	607,779	620,396	0.64	19.70	0.84	AGAP000043-RA
AGAP000070	protein kinase	X	1,098,557	1,107,555	0.87	21.10	0.88	AGAP000070-RB; AGAP000070-RA
AGAP000147	PDZ-binding kinase	X	2,329,924	2,334,904	0.12	15.90	0.74	AGAP000147-RA
AGAP000220	xylulokinase	X	4,089,326	4,094,219	0.85	21.20	0.85	AGAP000220-RA
AGAP000241	N-acetylglucosamine kinase	X	4,510,139	4,520,051	0.86			
AGAP000272	transcription initiation factor TFIID subunit 9 / adenylate kinase	X	4,994,393	4,996,244	0.90			
AGAP000310	mitogen-activated protein kinase kinase 3	X	5,468,098	5,470,361	0.73			
AGAP000338	phosphomevalonate kinase	X	5,884,575	5,885,561	0.61			
AGAP000368	PI-3-kinase-related kinase SMG-1	X	6,776,805	6,803,638	0.46			
AGAP000406	Rho-associated protein kinase	X	7,308,688	7,319,601	0.90			
AGAP000477	serine/threonine kinase 32	X	8,286,592	8,360,886	0.88			

Showing 1 to 15 of 262 entries (filtered from 13,094 total entries)

Copyright © 2022 Waterhouse Grou

Figure 3.3: online browsable table of gene-level metrics generated from the results of *mwgaw-analyses*. **A)** Table displaying the gene ID (from the input GFF), associated product, genomic location, phastCons score, alignability and identity scores, and transcripts for each gene in the reference annotation. **B)** CodAlignView visualisation of the input MWGA around the genomic region containing a transcript, accessed by clicking the hyperlink on the corresponding transcript ID in the main table.

Genome-wide sequence alignability in mosquitoes

Building reference MWGAs for available anopheline mosquito genome assemblies provides a rich nucleotide-level comparative data resource with which to explore the evolution of their genes and genomes. Using the alignability scores computed by the *mwgaw-analyses* workflow, a genome-wide visualisation of the extent and distributions of alignable sequences across multiple mosquito species' genomes provides a synthesis of the alignment information contained in the MWGA (**Figure 3.4**). Each reference genome assembly exhibits variable fractions of the total assembly that are alignable to all, some, or none of the others, with total alignable fractions ranging from 55% for *A. sinensis* to 82% for *A. darlingi*. A conserved core of about 14 Mbp is alignable across all 21 assemblies, representing 5.2% of the *A. gambiae* PEST genome. Most of the other alignable fractions vary for each assembly in a manner that reflects the species phylogeny, e.g. larger fractions of seven-species alignments for the seven closely-related members of the *A. gambiae* complex, and larger fractions of two-species alignments for the two members of each of the Nyssorhynchus and Anopheles subgenera, and the Neomyzomyia group. Much of the remaining unalignable sequence corresponds to masked repetitive sequences or assembly gaps. Anopheline chromosomes consist of five major elements, the X chromosome and the L and R arms of chromosomes two and three. These elements generally show a high level of conservation of gene content (macrosynteny) with large-scale rearrangements occurring via translocations of intact elements, unlike in *Drosophila* where fusions or fissions have occurred (Neafsey et al., 2015). Plotting the numbers of aligned species along the lengths of each major *A. gambiae* element averaged over 2-Kbp overlapping windows reveals dramatic regional differences in genome alignability. Small peaks of anopheline-wide alignable regions are distributed throughout each element, emerging from a plateau of less-alignable regions interspersed with valleys of poorly-alignable regions. Regions with high proportions of masked or gapped base pairs show an expected reduction in alignability, and are clearly evident at the starts of arms 2L and 3L and the ends of arms 2R and 3R and the X chromosome, which correspond to the locations of the centromeres. These chromoplots also highlight how the X chromosome sharply contrasts the autosomes with its generally lower levels of alignability, suggesting a faster rate of sequence evolution that leads to the loss of significantly recognisable homology. These visualisations demonstrate the genome-wide levels of and local variations in alignable genomic sequences using the alignability metric computed by running the *mwgaw-analyses* workflow on an input MWGA.

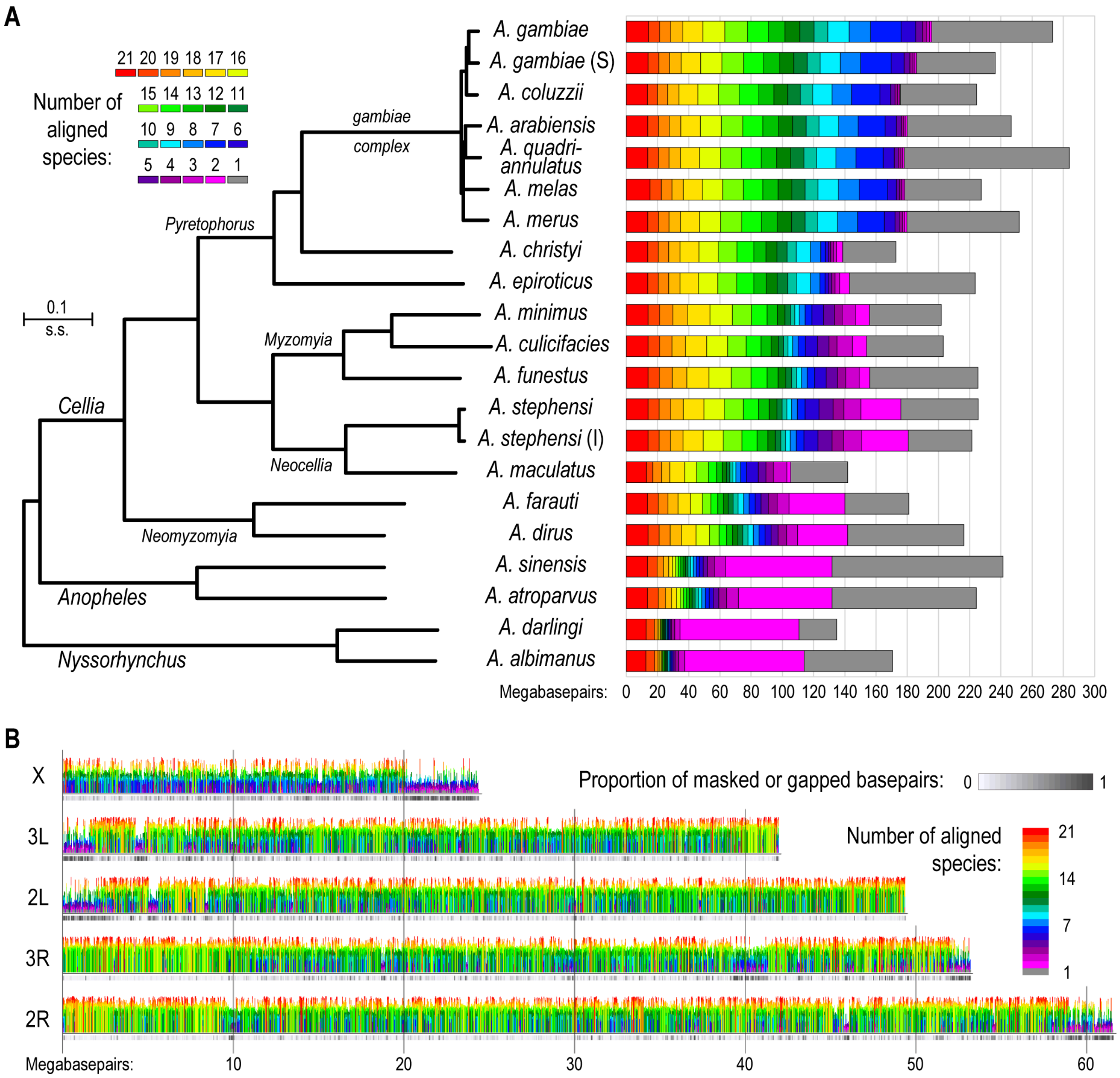


Figure 3.4: whole-genome multiple sequence alignments of 21 *Anopheles* genome assemblies. (A) The species phylogeny shows the relationships amongst the 21 anophelines with substitutions per site (s.s.) computed with the phyloFit tool from the PHAST analysis suite using four-fold degenerate sites of *Anopheles gambiae* protein-coding genes. The bars partition each genome assembly showing their alignability across the anophelines from 55% for *A. sinensis* to 82% for *A. darlingi*, with a conserved core of about 14 Mbps that is alignable across all 21 genomes. Approximately 180 Mbps are alignable for each of the seven members of the *A. gambiae* species complex, while for the remaining Cellia species this ranges from 105 Mbps for *A. maculatus* to 180 Mbps for *A. stephensi* (Indian). (B) The chromoplots show the genome alignability averaged over 2-Kbp windows along each of the five chromosomal arms for the *A. gambiae* genome assembly. Greyscale tracks beneath each plot show the proportion of masked or gapped base pairs over the same 2-Kbp windows. Densely masked or gapped regions exhibit reduced alignability, some of which correspond to the locations of centromeres at the starts of arms 2L and 3L and the ends of arms 2R and 3R and the X chromosome.

Discussion

Following the work presented in **Chapter 2** to build a reproducible workflow to compute MWGAs, this chapter describes a second reproducible workflow which efficiently implements multiple analyses quantifying patterns of sequence conservation from a MWGA. Indeed, the outcome of computing a MWGA is a large file in a format difficult to process without using specialised tools; the *mwgaw-analyses* transforms this fairly unusable and large file into 1) nucleotide-level resolution metrics, 2) gene-level metrics, and 3) visual representations of the information contained in the MWGA. Such metrics and visualisations are powerful tools to understand global patterns of sequence evolution contained in the alignment, as well as to explore the subtlety of local patterns in specific regions. To illustrate this point, we provide examples of analyses using the data generated for the 22 mosquito assemblies MWGA and the 36 *Drosophila* assemblies MWGA, showing how conservation of sequence expectedly matches the structure and content of the genome - for instance, capturing the higher rate of sequence evolution that is generally observed on the sex chromosomes in a heterogametic sex determining system (Beukeboom & Perrin, 2014). We also illustrate the usefulness of the two main visualisation outputs of *mwgaw-analyses*, genome browser track hubs and an interactive table of gene-level metrics, in exploring local patterns of sequence conservation - eventually leading to updating the existing annotation of *Anopheles gambiae* - and in identifying sets of genes of interest for gene-focused studies, which will be developed in **Chapter 4**.

In practice, the work presented in this chapter achieves two principal objectives: 1) through the *mwga-utils* software, which is further described in **Chapter 5**, it implements a framework for the computation of low-level metrics along a MWGA in the form of a MAF file, and 2) it greatly facilitates the computation of the most standard analyses of sequence conservation by encapsulating the usage of complex tools into a standardised and reproducible workflow. In both cases, the efforts expanded to implement the software and workflow in a documented fashion and following recommended practices for reproducible science greatly facilitate future additions to the workflow. In fact, both *mwga-utils* and *mwgaw-analyses* were designed to be extended, the former by implementing the computation of new metrics such as major allele frequency or allelic diversity per nucleotide, and the latter by adding additional analyses using MWGAs as the main input. Examples of such analyses include the *phyloFit* (Hubisz et al., 2011) and *phyloP* (Pollard et al., 2010) software, which were not implemented in the first version of *mwgaw-analyses* because of the redundancy of their results with that of phastCons, but could be easily added to the workflow to provide additional support to quantification of evolutionary constraints leading to sequence conservation.

In addition to providing a user-friendly, reproducible method of performing analyses on an MWGA, the *mwgaw-analyses* leverages the work invested in optimising runtime and peak memory usage of *mwgaw-align* to improve the scalability of compute-intensive analyses, namely the computation of metrics using phastCons and PhyloCSF. The latter in particular requires weeks of user runtime that were reduced to days thanks to the batching process implemented in *mwgaw-analyses*. One potential concern stemming from this batching process is that both PhyloCSF and phastCons rely on genome-level computations of estimates to adjust internal models prior to the scoring computation step; splitting the reference into batches and analysing these batches separately may therefore affect the value of estimates used in the models. However, the authors of each software both recommended a similar batching process in their user guides, and testing confirms that the impact of performing analyses on batches of sequences is negligible compared to estimating at the genome level. In the case of PhyloCSF, it is worth noting that a recent study attempted to optimise performance by releasing a novel efficient re-implementation of the underlying algorithm using a high-performance programming language (Pockrandt et al., 2022). We were involved in testing this new software and provided feedback during its development; while it does significantly decrease the runtime of the nucleotide-level estimation of coding potential, it is hindered by two limitations: first, it does not generate some intermediary files produced by the original implementation of PhyloCSF, which are necessary for the *pccr* analysis implemented in our workflow to generate a list of putative coding elements, and second, it does not allow the user to easily provide a custom model, making it effectively unusable to estimate coding potential for a MWGA comprising a set of assemblies for which existing models might be inappropriate. Because of these limitations, and although the performance increase is promising, this software was not included in *mwgaw-analyses* and the original implementation of PhyloCSF was used instead.

Going beyond the computation of nucleotide-level and gene-level metrics, the workflow presented in this chapter focuses on providing users with multiple visualisation outputs. When handling large amounts of data, as is often the case in genomics, having access to visualisation tools which can both summarise the data in a concise but comprehensive manner as well as enable exploration of the raw data at the genome level is invaluable. This paradigm has driven the work presented in this thesis, including also other software like *radsex*, presented in

Chapter 6, which also provides strong visualisation tools for genomic data. In the case of *mwgaw-analyses*, the efforts expanded into facilitating visualisation of data enable users to easily generate complex data packages in the form of genome browser track hubs, which can be integrated into existing public reference databases (UCSC GenomeBrowser, Ensembl, etc.) as well as private browser instance, including annotation platforms like the JBrowse-based Apollo (Dunn et al., 2019). Furthermore, the gene-level data already generated by the workflow could form the basis of additional visualisations integrated in the web-based browsable table. In particular, we developed scripts to generate gene-level figures displaying MWGA metrics for all genes in a provided annotation, which are used in our study of evolutionary constraints on immune gene families presented in **Chapter 4**. Although these tools are not yet mature enough to be directly plugged into *mwgaw-analyses*, their integration in both *mwgaw-analyses* and into the browsable table would require relatively little additional efforts thanks to the modular structure of the workflow and the automated generation of the browsable table website.

Chapter 4: Applying the MWGA workflows to explore functional constraints in immunity genes

Summary

This chapter serves to demonstrate the utility of the results generated by the workflows presented in **Chapter 2** and **Chapter 3** to explore the relationships between sequence conservation and gene functions in the biological context of the mosquito immune system. The mosquito multispecies whole genome alignment is used to estimate gene-level sequence conservation measuring long-term evolutionary constraint for *Anopheles gambiae* genes. Taking advantage of available population polymorphism data for this species, we also estimated gene-level sequence conservation measuring modern-day diversity across sampled populations from Africa. We use these two metrics to investigate the functional features of the most and least conserved mosquito genes, as well as examining constraint and diversity levels for functional categories of genes defined by their co-expression patterns. Comparing patterns of evolutionary constraint amongst different functional categories of immune-related genes showed that the ensemble of immunity genes contains both conservatively and dynamically evolving components, and that these distinctions appear to be linked to their functional roles. These results provide initial insights into how sequence conservation levels can vary according to a gene's functional role and illustrate how the MWGA workflows enable the reliable calculation of sequence conservation to explore genome-wide variation in evolutionary constraint and how this relates to gene function.

Introduction

Most of the 3,500 identified mosquito species are harmless to humans (Ruzzante et al., 2019), but several species act as vectors of some of the diseases causing the most human deaths around the world, notably malaria, which is caused by parasites of the genus *Plasmodium* transmitted by *Anopheles* mosquitoes, Dengue and yellow fever which are viral diseases propagated by *Aedes aegypti*, and West Nile virus which is transmitted by *Culex* mosquitoes. Because of their crucial role in public health issues, disease-vector mosquitoes have attracted a lot of research interest and have been prioritised for genomic studies among insects: the genome of *Anopheles gambiae* was the second insect genome to be sequenced (2002) after

Drosophila melanogaster, and *Aedes aegypti* was sequenced in 2007. Today, more than 30 genome assemblies are available for *Anopheles*, *Aedes*, and *Culex* mosquito species (Feron & Waterhouse, 2022a). These genomes constitute a powerful resource which can be leveraged to provide an evolutionary perspective on our biological understanding of disease-vector insects, in particular the factors involved in their interactions with pathogens that affect transmission, which are crucial for controlling the spread of debilitating diseases (Neafsey et al., 2015). Understanding the evolution and functions of these factors will help to develop novel approaches to limit the damaging effects of mosquitoes on human health by facilitating targeted interventions while minimising ecological knock-on effects.

Many of the interactions between mosquitoes and the pathogens they carry involve components from the innate immune system (Christophides et al., 2002). Like most arthropods, mosquitoes deploy a strong immune response when facing infection by a pathogen, which makes them remarkably resistant to these infections. This response incorporates multiple molecular components to recognize pathogens, transfer the recognition signal, and activate and modulate the pathogen-killing molecules. Examples of pathogen-recognition receptors include peptidoglycan recognition proteins (PGRPs) (Q. Wang et al., 2019) and β-1,3-glucan recognition or gram-negative bacteria-binding proteins (GNBPs) (Rao et al., 2018). Signals from these receptors are transmitted by cascades like the Toll (Valanne et al., 2011) and the JAK/STAT (Myllymäki & Rämet, 2014) pathways to regulate the expression of effector genes, for instance genes encoding antimicrobial peptides (AMPs) (Lazzaro et al., 2020). These different components interact through a diversity of cellular processes controlled by cells from multiple tissues to form a complex network of interactions, which provide mosquitoes with a strong protection against the majority of pathogens.

The overall structure of the immune system and the gene families involved in the immune response are shared across insects. However, the molecular components of the immune system are engaged in an arms race with pathogens and therefore the genes encoding these components are expected to be under strong selective pressure and thus evolve faster than the majority of genes involved in other biological processes (Obbard et al., 2009). In practice, genes involved in different processes of the immune response are affected differently by this arms race, and the genes encoding components interacting directly with pathogens, for instance recognition proteins, are expected to evolve the fastest. Understanding precisely which immune genes are under which evolutionary constraints is key to identifying the components of the mosquito immune system that are involved in this arms race with pathogens; when mosquitoes lose this race, they become vectors of the deadly diseases for which they are known (Bartholomay & Michel, 2018). This outcome was the driver of early studies comparing immune genes in *Anopheles gambiae* and *Drosophila melanogaster* to identify mosquito immune genes under selection by looking at gains and losses for different families (Christophides et al., 2002). Later studies targeting *Anopheles gambiae* and *Aedes aegypti* expanded on these results by comparing the sequences of immune genes in 11 insect species, focusing on conserved domains (Waterhouse et al., 2007). Since then, more than 25 mosquito species have been sequenced, which allows us to further refine the analysis of evolutionary conservation of immune genes in mosquitoes by estimating evolutionary constraints from sequence conservation computed using multispecies whole genome alignments (MWGAs). In parallel, sequencing projects under the umbrella of the Malaria Genomic Epidemiology Network (MalariaGEN) now provide population-level genomic diversity datasets in the form of Single Nucleotide Polymorphism (SNP) for almost 2,800 *Anopheles gambiae* individuals (The Anopheles gambiae 1000 Genomes Consortium, 2021). Thanks to these data, it is now possible to compute estimates of sequence conservation both at the lineage level and at the population level, and thus compare long-term against short-term evolutionary constraints.

In this chapter, we use the MWGA of the genome sequence of 22 mosquito species that we described in **Chapter 2** to compute gene-level sequence conservation estimates using *mwgaw-analyses*, described in **Chapter 3**, for all genes in the annotation of the reference assembly for *Anopheles gambiae*. We then compute an estimate of population-level sequence conservation using the MalariaGEN SNP data for each of these genes. We first use these data to survey the landscape of sequence conservation in the entire gene set, identifying most and least conserved genes in the annotation, and exploring the biological functions associated with these genes. We then dive into specific sequence conservation patterns for different immune gene families and investigate how these patterns relate to the functional roles of these genes in the mosquito immune system.

Methods

Estimating gene sequence conservation from the 22 mosquito MWGA

To compute estimates of conservation of sequence over evolutionary times, we used the 22 mosquito MWGA described in **Chapter 2**. Briefly, this alignment comprises assemblies for 22 mosquito species: one *Culex* species, two *Aedes* species, and 19 *Anopheles* species, including the one whose assembly was used as reference, *Anopheles gambiae*; the divergence time between these species is estimated to be ~ 150 Mya. The resulting MWGA was used as the primary input for *mwgaw-analyses*, described in **Chapter 3**, along with the phylogenetic tree used to generate the MWGA with *mwgaw-align*, and a GFF file containing the annotation for *Anopheles gambiae* (AgamP4, annotation version 4.11). The workflow was run on UNIL's high performance computation platform using the following phastCons parameters: a target coverage of conserved elements of 0.3 (--target-coverage 0.3) and an expected length of conserved elements of 35 (--expected-length 35). The workflow generated an average score of sequence conservation for each gene in the provided annotation, exported as a tabulated file.

Estimating gene sequence conservation from population sampling data

The MalariaGEN project provides SNP data for thousands of *Anopheles gambiae* individuals and is updated with additional data over time. At the time of this study, data were available for 1,470 individuals from 28 different sample sets, with 10 to 303 individuals per sample set. For each sample, metadata were first downloaded from a google cloud bucket using *gsutil* version 5.10 in *rsync* mode (-m rsync) from the google cloud bucket provided by MalariaGEN (gs://vo_agam_release/v3/metadata/). The resulting metadata are organised into cohorts reflecting the sampling process; these cohorts can contain individuals belonging to species other than *Anopheles gambiae*. Sample metadata files were processed using a Python script for each cohort in order to extract *Anopheles gambiae* samples. Then, a Variant Calling Format (VCF) file containing individual SNP information was downloaded from google cloud for each identified sample using *wget*, and immediately compressed with *bcftools view* using the following parameter values to retain only polymorphic sites: *--output-type z --min-ac 1:nonmajor --trim-alt-alleles*. Sample information that was not needed for this analysis was removed from each VCF using a custom script in order to reduce file size. The resulting compressed VCF were indexed with *bcftools index* and merged into a single, multi-sample VCF using *bcftools merge*. This process was implemented as a small Snakemake workflow executed on UNIL's HPC platform and resulted in a single VCF containing all polymorphic sites for all *Anopheles gambiae* samples in the MalariaGEN dataset.

Variants in the merged VCF file were annotated with *SnpEff* version 5.1d (Cingolani et al., 2012), using the pre-existing database for the *Anopheles gambiae* genome. This software creates a new VCF with an added annotation column containing pre-computed information on each polymorphic site, as well as a summary table of association between each site and a set of tags related to the estimated effect of the associated variant. The summary table was parsed using a custom Python script to classify each polymorphic site into a synonymous or non-synonymous variant based on the tags annotated by *SnpEff*. Then, another script was used to count the number of synonymous and non-synonymous variants for each gene in the reference *Anopheles gambiae* annotation, exported as a tabulated file. The counts of synonymous vs non-synonymous sites in a gene were used to compute the following estimate of population-level sequence conservation for this gene: **NS / (NS + S)**, with **NS** being the number of non-synonymous SNPs and **S** the number of synonymous SNPs; the value of this estimate ranges from 0 when all SNPs in the focal gene are synonymous, to 1 when all SNPs are non-synonymous. Therefore, a low value of this estimate suggests that most of the variable sites in the focal gene are under constraint to encode a fixed amino acid, and thus the resulting protein sequence is relatively conserved; conversely, a high value of this estimate suggests that variable sites in the focal genes are not constrained to encode a fixed amino acid, and thus the resulting protein sequence is variable in the population. To reliably compute this estimate, genes with fewer than ten polymorphic sites were discarded from the analysis.

Visualising the metrics computed for all genes

To fully leverage the data generated for this study, we implemented a suite of scripts and functions using R (R Core Team, 2024). Using these tools, we were able to automatically generate two novel visualisations: 1) “kite” plots showing the distribution of both population-level and lineage-level sequence conservation for all genes in the reference annotation, and 2) “track” plots showing all the computed information for each individual gene in the reference annotation.

Distribution of sequence conservation amongst genes in the reference annotation

In this study, we computed estimates of sequence conservation at the gene level at two different scales: across species using the 22 mosquito MWGA, and within the *Anopheles gambiae* population using SNP data. To visualise the distribution of these two metrics for each gene in the reference annotation, we implemented a series of R scripts which take as input a table of each metric for each gene in the annotation and one or more sets of genes to highlight. For each user-provided set of genes to highlight, the median, first quartile, and third quartile values of each of the two metrics are computed. The results are displayed in a figure on which the horizontal axis represents the population-level sequence conservation estimate **NS / (NS + S)** for a gene as described in the previous section, the vertical axis represents the phastCons score for a gene, and each gene is represented by a dot at its respective coordinates, with the given set of genes highlighted in user-controlled colours. For each of the highlighted sets of the genes, the median, first quartile, and third quartile are drawn for both metrics and connected to form a losange resembling a kite. This “kite” is akin to a two-dimensional box plot, with its centre indicating the median of both values, and each vertex indicating one of the quartile values. Examples of kite plots are shown in **Figures 4.5** and **4.7**.

Visualising computed metrics along the length of a gene

In the context of this study, we computed multiple gene-level metrics for all genes in the *Anopheles gambiae* genome annotation dataset. In order to better understand how these metrics relate to the gene sequence and structure, and to explore local variations along the sequence of a gene, we needed a way to automatically display all this information for any chosen gene. To this aim, we developed a suite of R scripts and functions which allowed us to programmatically generate a figure including multiple information tracks for any user-supplied gene in our dataset. An example showing all possible information tracks for the gene AGAP009263 (a member of the CLIP immune gene family) is presented in **Figure 4.1**. From top to bottom, the first track displays the gene model, including exons, untranslated regions (UTRs), coding sequences (CDSs), and known protein domains annotated by PFAM (El-Gebali et al., 2019); this model is drawn with custom functions using a tabulated input file containing information on each part of the gene. The second part displays genomic regions identified as repeated sequences, obtained from a repeat GFF file for the reference assembly. The third track shows the position of annotated variable sites identified from the SNP data, with variants annotated as synonymous coloured in blue and variants annotated as non-synonymous coloured in red. The fourth track shows the density of SNPs in a sliding window, including all sites from the original SNP data. Finally, the fifth track displays the Alignability (in orange) and phastCons score (in blue) per nucleotide computed with *mwgaw-analyses* for each nucleotide along the sequence of the gene.

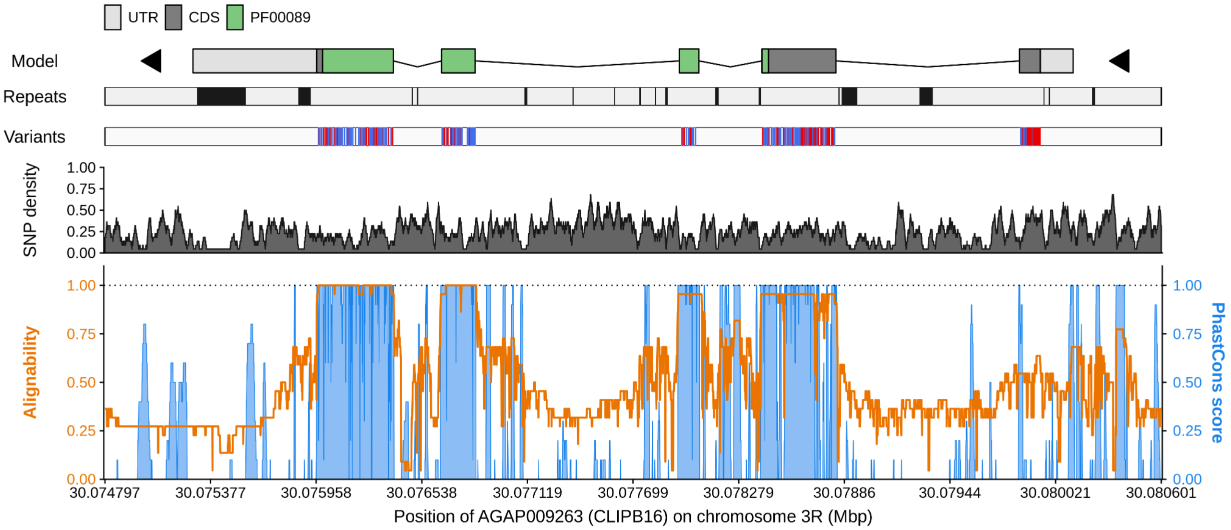


Figure 4.1: individual gene plot with all possible information tracks for the gene AGAP009263. From top to bottom, the tracks show: 1) the gene model, including exons, untranslated regions (UTRs), coding sequences (CDSs), and known protein domains annotated by PFAM; 2) genomic regions identified as repeated sequences; 3) position of annotated variable sites identified from the SNP data, with variants annotated as synonymous coloured in blue and variants annotated as non-synonymous coloured in red; 4) density of all detected SNPs in the original data in a sliding window; 5) alignability (in orange) and phastCons score (in blue) per nucleotide along the sequence of the AGAP009263.

Results

Functional features of the most and least constrained mosquito genes

The 22 mosquito MWGA was used to compute genome-wide evolutionary conservation scores with phastCons and the *Anopheles gambiae* polymorphism data from the MalariaGEN project were used to qualify synonymous and non-synonymous SNPs in protein-coding genes. Per-nucleotide phastCons scores were averaged along the coding sequences to estimate per-gene constraint levels, and the proportion of non-synonymous SNPs out of all SNPs within each gene was calculated to estimate per-gene diversity levels. The distributions of constraint and diversity levels measured for ~12,000 *Anopheles gambiae* protein-coding genes show a median of 0.57 for diversity and 0.71 for constraint (**Figure 4.2**). Diversity levels are relatively symmetrically distributed with a fairly narrow interquartile range and very few extreme values. Constraint levels are skewed towards higher values with a wider interquartile range and substantial numbers of genes across the entire spectrum. Visualising these distributions demonstrates that genes exhibit a wide range of conservation levels as measured by long-term evolutionary constraint and population-level diversity.

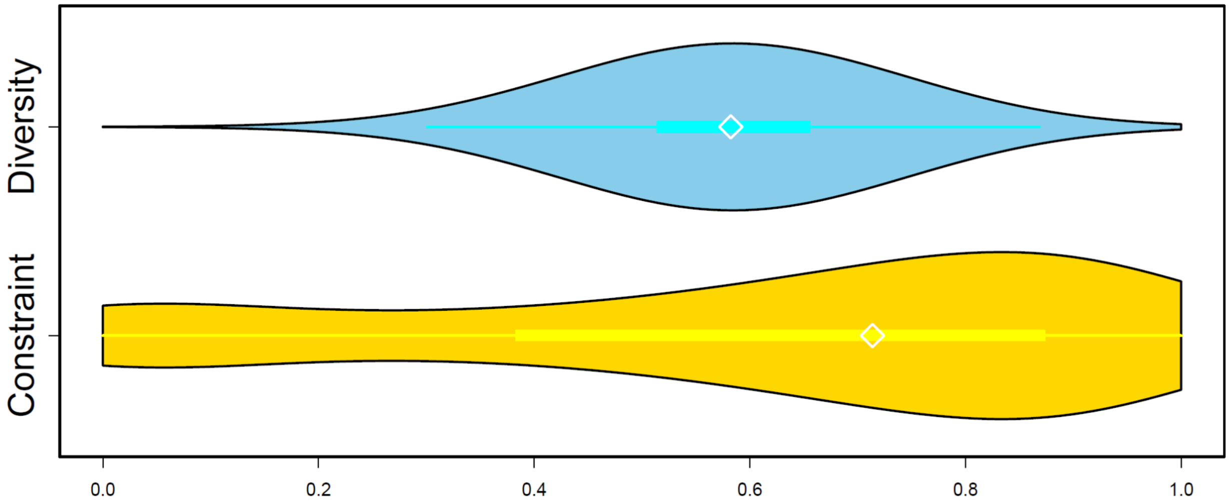


Figure 4.2: geneset-wide distributions of computed constraint (phastCons) and diversity (proportion of non-synonymous SNPs) metrics for ~12,000 *Anopheles gambiae* protein-coding genes. The distributions span the ranges of low to high diversity and low to high constraint, with a more symmetrical distribution for diversity and a more skewed distribution for constraint. The violin plots show kernel density estimates of the distributions, and the boxplots within indicate the medians (white-border diamonds), the interquartile range (thick lines) and 1.5 x the interquartile range (thin lines). The values on the x-axis range from 0 to 1, the permissible values for the computed constraint and diversity values.

To begin to investigate broad relationships between sequence conservation and gene function, Gene Ontology (GO) enrichment analyses using the Kolmogorow-Smirnow test were performed with TopGO (Alexa & Rahnenfuhrer, 2024) and visualised using GO-Figure! (Reijnders & Waterhouse, 2021). Results for enriched GO Biological Processes from the phastCons constraint levels for the most constrained and least constrained genes are shown in **Figure 4.3** and **Figure 4.4**, respectively. For the most constrained genes, the analysis identifies several core housekeeping functions such as translation, transcription, splicing, mitosis, protein transport and secretion, neuronal processes, development and morphogenesis, and phagocytosis. For the least constrained genes, the analysis identifies several functions implicated in potentially dynamically evolving processes such as taste and chemical stimulus perception, defence responses to viruses, chitin metabolism, GPCR signalling, and courtship behaviour. The equivalent analyses for gene diversity levels also identified core housekeeping functions enriched amongst low-diversity genes, such as transcription, splicing, mitosis, protein transport, and neuronal processes. In contrast, amongst high-diversity genes enriched processes included taste and chemical stimulus perception, chitin metabolism, and GPCR/iGluR signalling, which, like for low-constraint genes, are functions implicated in potentially dynamically evolving processes. This global analysis of protein-coding genes demonstrates that these conservation metrics correctly characterise genes expected to be constrained by their key biological functions with low-diversity and high-constraint levels, and conversely genes with functions associated with more rapidly evolving processes with high-diversity and low-constraint levels.

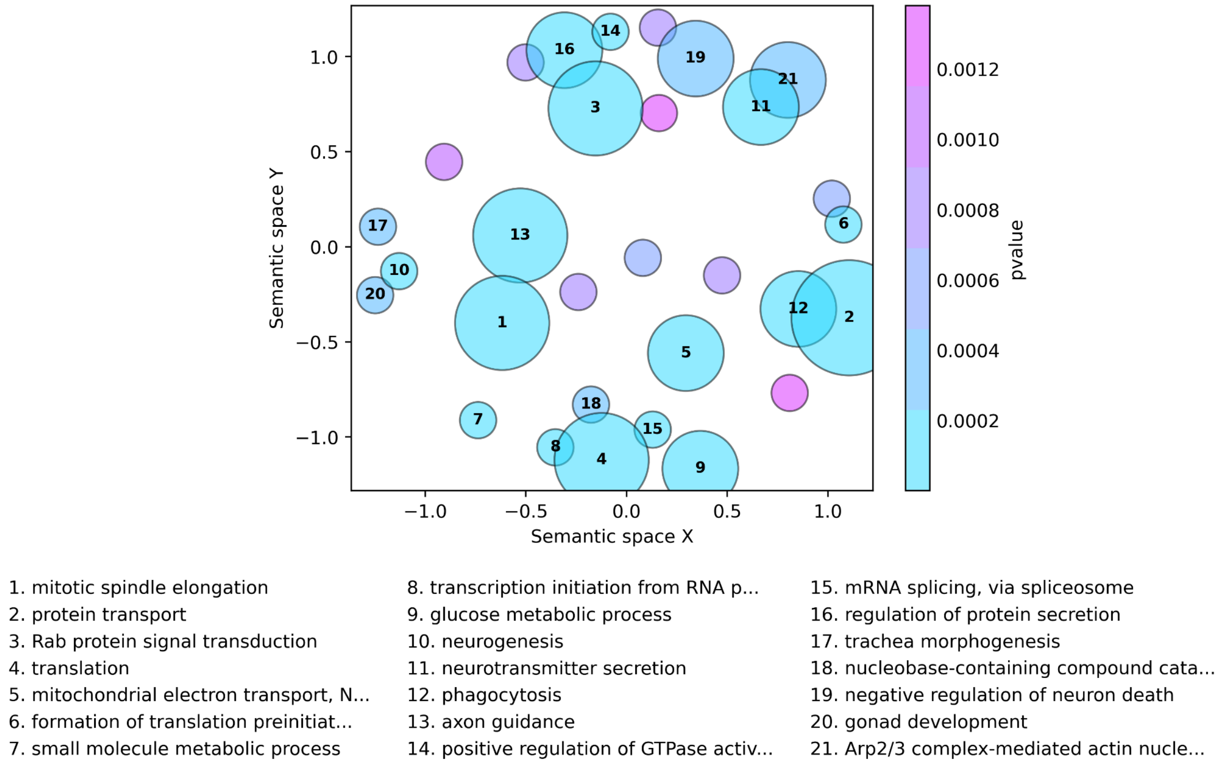


Figure 4.3: summary visualisation of Gene Ontology (GO) enrichment amongst the most constrained genes (highest phastCons scores). Enrichment analysis identifies several core housekeeping functions such as translation, transcription, splicing, mitosis, protein transport and secretion, neuronal processes, development and morphogenesis, and phagocytosis. Each bubble groups one or more semantically related GO Biological Process terms, the bubbles are arranged such that those that are most similar in semantic space X and Y are placed nearest to each other, they are coloured according to the enrichment p-value of the representative term for each bubble. The GO term names are shown for the top 21 most significantly enriched terms. The figure was produced using GO-Figure! (Reijnders & Waterhouse, 2021) and the enrichment analysis using the Kolmogorow-Smirnow test was performed with TopGO (Alexa & Rahnenfuhrer, 2024).

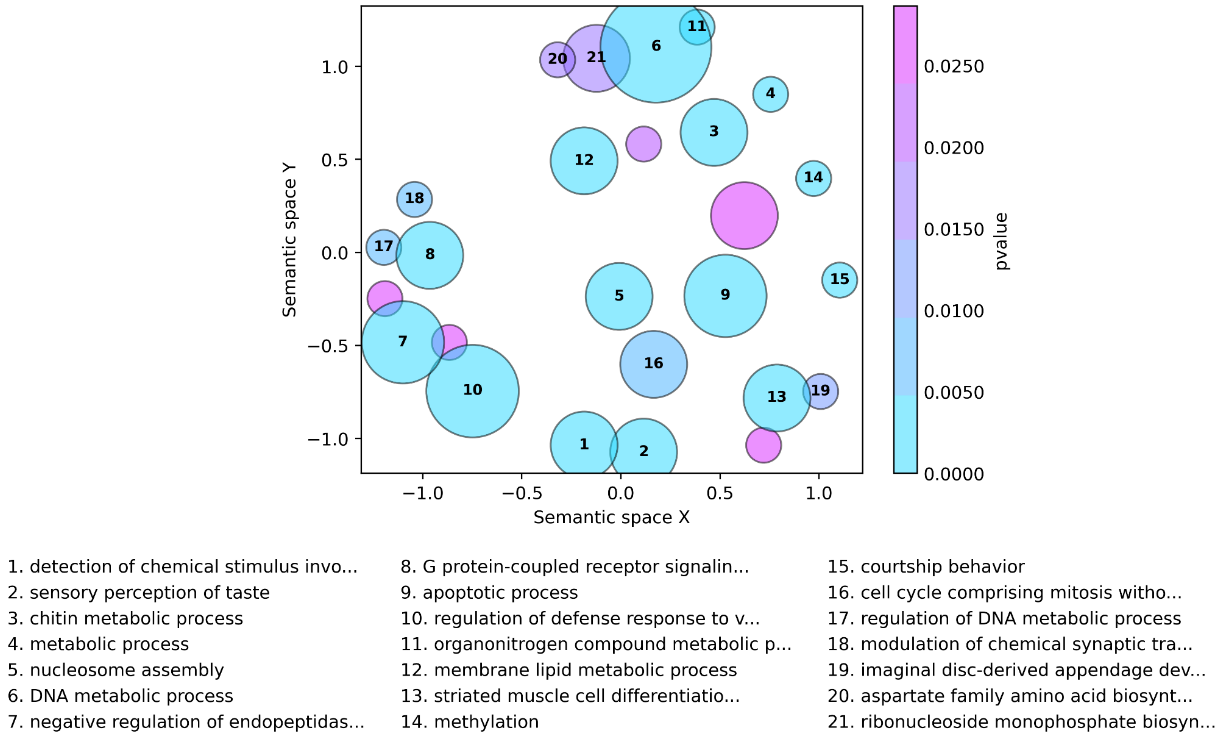


Figure 4.4: summary visualisation of Gene Ontology (GO) enrichment amongst the least constrained genes (lowest phastCons scores). Enrichment analysis identifies several functions implicated in potentially dynamically evolving processes such as taste and chemical stimulus perception, defence responses to viruses, chitin metabolism, and courtship behaviour. Each bubble groups one or more semantically related GO Biological Process terms, the bubbles are arranged such that those that are most similar in semantic space X and Y are placed nearest to each other, they are coloured according to the enrichment p-value of the representative term for each bubble. The GO term names are shown for the top 21 most significantly enriched terms. The figure was produced using GO-Figure! (Reijnders & Waterhouse, 2021) and the enrichment analysis using the Kolmogorow-Smirnow test was performed with TopGO (Alexa & Rahnenfuhrer, 2024).

Gene expression data can provide an orthogonal approach to gene functional categorisation that does not rely entirely on Gene Ontology annotations. Therefore, to further explore broad relationships between sequence conservation and gene function, the *Anopheles gambiae* gene co-expression network built by (Kuang et al., 2022) was used to define and compare clusters of co-regulated genes putatively involved in common biological processes. This network included 15 “communities” of co-regulated genes, where each community was enriched for distinct GO biological processes ranging from fundamental cell functions to specialised physiological roles associated with specific organs or tissues. For example, Community 4 was associated with intracellular signal transduction, protein phosphorylation, and neuron function, while Community

7 was associated with innate immunity and lipid metabolism, and Community 15 was associated with blood feeding and salivary glands. The network core, characterised by strong gene expression across the majority of conditions, was enriched in genes required for core housekeeping processes of oxidative phosphorylation and translation. Visualising the sequence conservation levels of these sets of putatively functionally related genes as “kite plots” (see Methods above) clearly demonstrates how different functional categories can show dramatically contrasting patterns of conservation (**Figure 4.5**). The network core genes expectedly exhibit consistently high levels of evolutionary constraint and reduced levels of diversity compared to the background of all other *Anopheles gambiae* genes. In stark contrast, Community 15 “blood feeding and salivary gland” genes exhibit dramatically reduced levels of evolutionary constraint and generally above average levels of population diversity. Mosquito salivary gland genes are known to be rapidly and dynamically evolving (Arcà et al., 2017), demonstrating that these conservation metrics correctly characterise genes expected to be associated with more rapidly evolving processes with high-diversity and low-constraint levels.

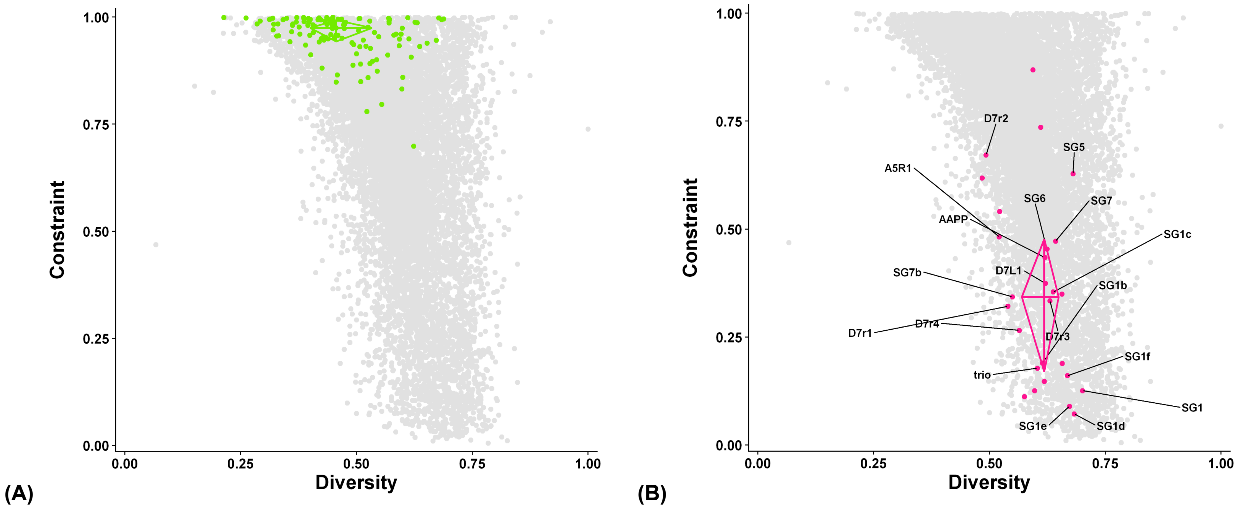


Figure 4.5: kite plots showing evolutionary constraint and population-level diversity measures of *Anopheles gambiae* genes. **(A)** Network core genes, as defined by the co-expression network analysis of (Kuang et al., 2022), enriched in genes required for core housekeeping processes of oxidative phosphorylation and translation, are plotted in green with all other genes shown in grey. **(B)** Community 15 “blood feeding and salivary gland” genes, as defined by the co-expression network analysis of (Kuang et al., 2022), are plotted in pink with all other genes shown in grey. The “kites” correspond to two-dimensional boxplots, with the centre indicating the median of both values (constraint and diversity), and each vertex indicating one of the quartile values.

Conservation and divergence of mosquito immune-related genes

Immunity gene functional categories display distinct evolutionary constraints

Exploring the relationships between sequence conservation and gene function in the context of the insect innate immune system has revealed distinct evolutionary dynamics that characterise different components (Bartholomay et al., 2010; Waterhouse et al., 2007, 2010). These and other studies developed the framework of classifying immune system components according to their functional roles, including in pathogen recognition processes, immune signalling pathways, modulation of response activation or deactivation, antimicrobial activities, autophagy and RNA interference. Using this framework, together with evolutionary constraint levels computed using phastCons on the mosquito multispecies whole genome alignment data, we could compare and contrast levels of sequence conservation amongst different immunity gene functional categories (**Figure 4.6**). The immune system is traditionally thought of as rapidly and dynamically evolving, but this analysis shows that while some functional components do indeed display low constraint levels, several other functional categories are in fact characterised by genes with much higher levels of sequence conservation.

Classical recognition gene families such as galectins (GALEs), Gram-negative binding proteins (GNBPs), peptidoglycan recognition proteins (PGRPs), and scavenger receptors (SCRA, SCRB) are relatively constrained despite interacting with pathogens. This may reflect the relatively limited structural diversity of the main microbial ligands - peptidoglycan, β-1,3-glucan, lipoproteins - they bind to or cleave (Ruzzante et al., 2022). Other recognition gene families exhibit much lower levels of constraint, especially C-type lectins (CTLs), fibrinogen-related proteins (FREPs), leucine-rich repeat immune proteins (LRIMs), and thioester-containing proteins (TEPs). In the case of MD-2-like proteins (MLs), their conservation more closely resembles classical recognition families, possibly warranting their reclassification, especially considering that they can bind lipopolysaccharides from the outer membrane of Gram-negative bacteria (Ruzzante et al., 2022).

Components of the three main signalling pathways - the immune deficiency pathway (IMD), janus kinase protein (JAK)/signal transducer and activator of transcription (STAT) pathway, and the Toll pathway are usually subdivided into signalling proteins (SIG) and modulator (MOD) proteins. The constraint analysis shows that while the sequences of genes that function as modulators are generally highly conserved, signalling protein sequences are generally less constrained. The conservative sequence evolution of modulators may be linked to their functional roles as enzymes, such as ubiquitinases like Effete and Bendless, or E3 ligases like Pellino and Pias, where maintaining their enzymatic activities means only low sequence divergence can be tolerated (Ruzzante et al., 2022).

Almost all families of cascade modulators, which comprise caspases (CASPs), CLIP-domain serine proteases (CLIPA/B/C/D/E), inhibitors of apoptosis (IAPs), and serine protease inhibitors (SRPNs), show low sequence conservation levels (in this analysis only CLIPBs were significantly lower than average). Knowledge of the molecular functions of genes from these families is limited, nevertheless, it appears that functional modulator modules regularly re-source components from relatively large gene families, which could explain why long-term evolutionary constraint is weak (Ruzzante et al., 2022).

Amongst the other families analysed, antimicrobial peptides (AMPs) exhibit an average level of sequence conservation, while lysozymes (LYSs) are much less constrained. The effector enzymes including glutathione, heme, and thioredoxin peroxidases (GPXs, HPXs, TPXs) as well as superoxide dismutases (SODs), show high levels of sequence conservation, consistent with their molecular functions as enzymes controlling the production of reactive oxygen species and other immune defence biomolecules. Autophagy-related (APHAG) and small regulatory RNA pathway (SRRP) members are also more conserved than average, consistent with both autophagy and RNAi being ancient cellular processes with roles beyond immunity (Ruzzante et al., 2022). The Toll receptors (TOLLs) show generally elevated levels of sequence conservation, which could be consistent with insect TOLLs being activated by cytokines (e.g. spaetzle-like proteins, SPZs) rather than by binding pathogens as in the case of vertebrate Toll-like receptors, where patterns of positive selection are concentrated in ligand-binding domains and suggest host-pathogen coevolutionary interactions (Liu et al., 2020).

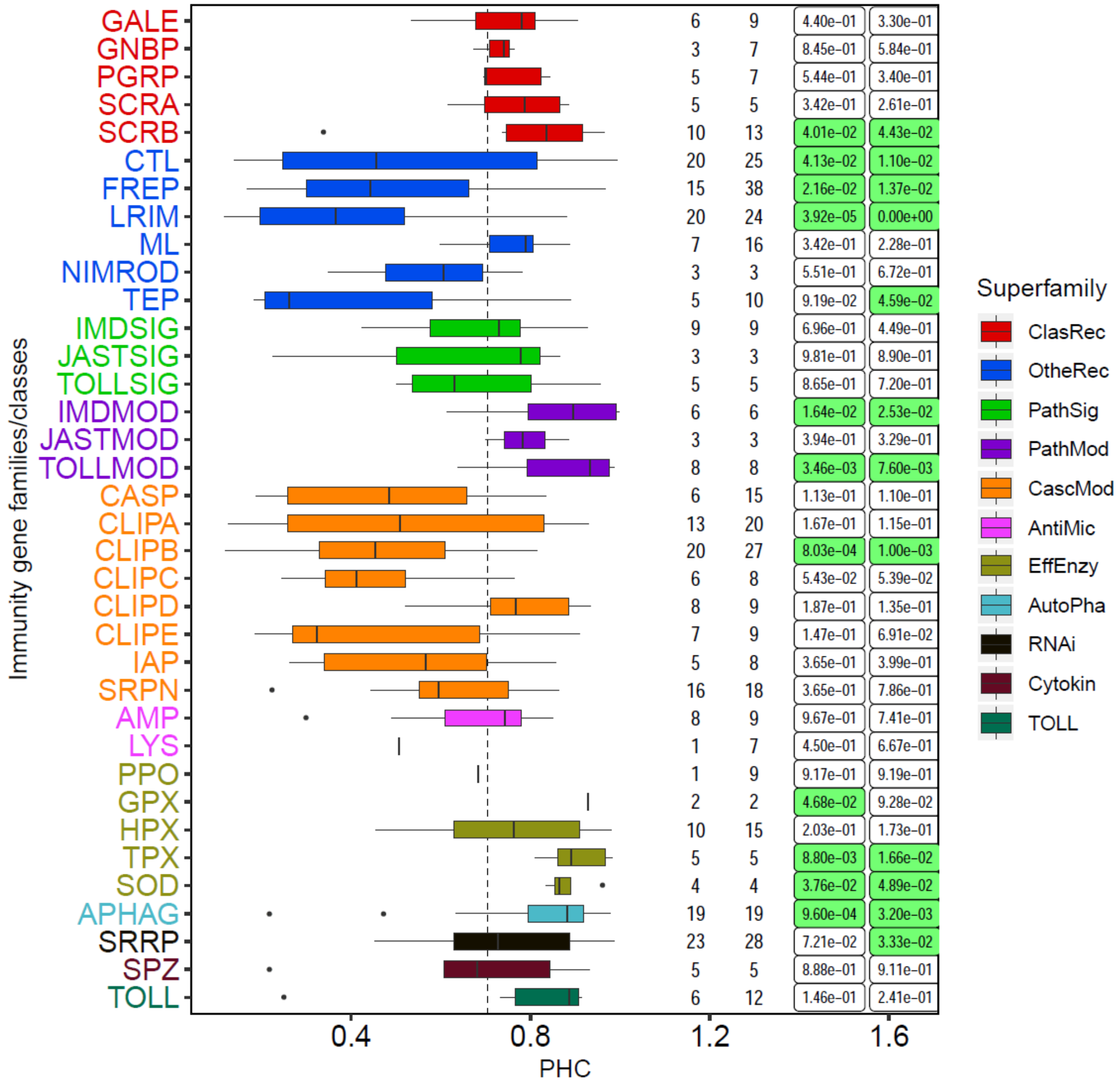


Figure 4.6: distributions of computed Orthologous Group (OG) metrics for all of the immune gene families for their phastCons (PHC) scores, with statistical assessments of the significance of deviations from the typical values across all families. Per family data, coloured by superfamily: OGs, number of orthologous groups; Genes, number of genes; MW p-val, Mann Whitney U test p-value; PRM p-val, permutation test p-value. P-values less than 0.1 are highlighted. Immune gene superfamilies: ClasRec, classical recognition; OtheRec, other recognition; PathSig, pathway signalling; PathMod, pathway modulation; CascMod, cascade modulation; AntiMic, antimicrobial peptides; EffEnzy, effector enzymes; AutoPha, autophagy genes; RNAi, RNA interference; Cytokin, cytokines; TOLL, Toll receptors. Reproduced from Additional File 1 from (Ruzzante et al., 2022).

Constraint and diversity within and between immune gene families

The computation of per-gene measures of evolutionary constraint and population diversity allows for contrasting long-term sequence conservation patterns with current polymorphism levels of individual genes and gene families. This is demonstrated here (**Figure 4.7**) by examining mosquito antimicrobial peptides (AMPs) - attacin (ATT), cecropins (CECs), defensins (DEFs), and gambicin (GAM); the classical recognition proteins - Gram-negative binding proteins (GNBPs) and peptidoglycan recognition proteins (PGRPs); and the mosquito-specific leucine-rich repeat immune proteins (LRIMs). Taken together, AMPs are not substantially different from the background of all other *Anopheles gambiae* genes in terms of their median constraint and diversity levels. Nevertheless, there is a large variation in constraint levels between the more conserved *CEC2*, *CEC3*, and *GAM1*, and the very low constraint shown by *DEF5*. The low sequence conservation levels of *DEF5* meant that in early work on immunity in *Anopheles gambiae* this gene was completely missed by annotation pipelines and therefore not included in the evolutionary analyses (Waterhouse et al., 2007). It is evolutionarily young, appearing after the split between old and new-world mosquitoes some 100 million years ago, most likely originating from a retroposition event of *DEF4*. The classical recognition proteins present a similar distribution to the AMPs in terms of constraint and diversity, with overall slightly higher conservation levels and no extremely low-conservation members like the *DEF5* AMP. *PGRPS1* and *PGRPS2* are the most highly conserved of the PGRPs, a situation that might be impacted by the potential for gene conversion occurring between these two neighbouring genes, as has been reported for *Drosophila* PGRPs (Jiggins & Hurst, 2003). While some LRIMs are amongst the most highly constrained and least diverse genes (e.g. *LRIM3*, *LRIM19*), overall there is a striking contrast between the LRIMs and the classical recognition proteins with the LRIMs showing dramatically reduced levels of sequence constraint and generally elevated levels of diversity. LRIMs are a mosquito-specific gene family with several members having been implicated in immune complement cascades that are important especially for susceptibility to *Plasmodium* infection (Waterhouse et al., 2010). Some LRIMs have been extensively studied and exceptional population-level diversity has been reported e.g. for *LRIM1* and *ALP1A*, *ALP1B*, and *ALP1C* (Holm et al., 2012; Rottschaefer et al., 2011), in agreement with the results from our constraint and diversity analyses. *LRIM9*, which also exhibits low constraint and high diversity, was not amongst the first studied genes, but when tested experimentally it also exhibited a key role as a *Plasmodium berghei* antagonist with phenotypes distinct from family members *LRIM1* and *APL1C* (Upton et al., 2015). This suggests that further studies of other low-constraint high-diversity LRIMs, like *LRIM16A*, *LRIM16B*, *LRIM8A*, *LRIM8B*, or *LRIM26* might prove fruitful avenues to dissect mechanisms of pathogen-mosquito interactions, where for example recent single-cell transcriptomics have identified *LRIM26* as a marker gene for immune cells known as granulocytes (Kwon et al., 2021).

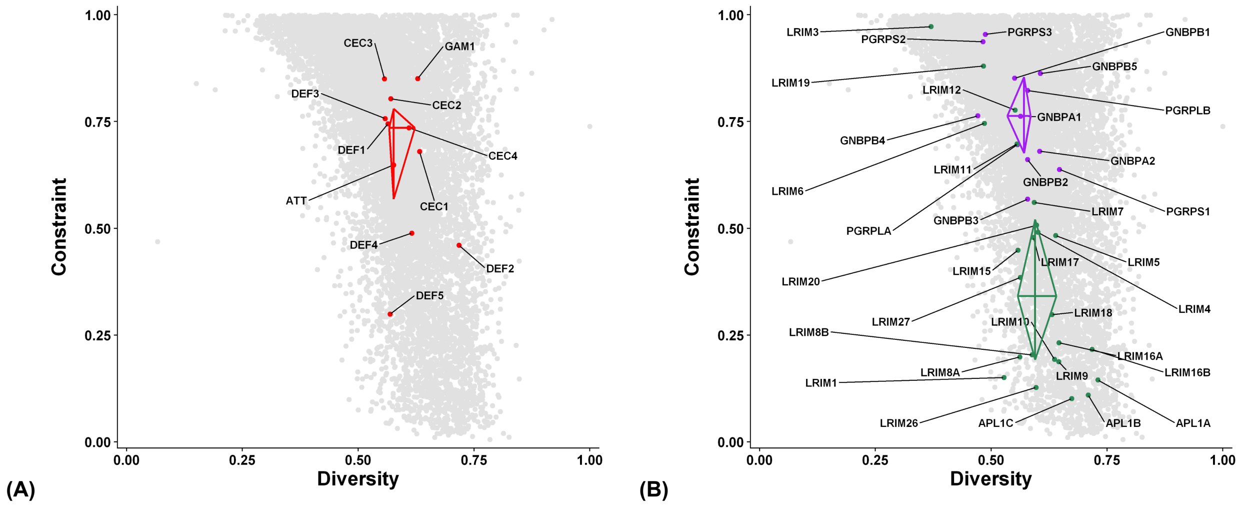


Figure 4.7: kite plots showing evolutionary constraint and population-level diversity measures of *Anopheles gambiae* immunity genes and gene families. The “kites” correspond to two-dimensional boxplots, with the centre indicating the median of both values (constraint and diversity), and each vertex indicating one of the quartile values. The named genes shown on the plots belong to families of antimicrobial peptides **(A)**: attacin (ATT), cecropins (CECs), defensins (DEFs), and gambicin (GAM); and **(B)** the classical recognition proteins (purple): Gram-negative binding proteins (GNBPs) and peptidoglycan recognition proteins (PGRPs); and the mosquito-specific leucine-rich repeat immune proteins (LRIMs, green).

Discussion

To explore relationships between gene conservation and gene function, we first examined the functional features of the most and least constrained mosquito genes using conservation scores estimated with phastCons from whole genome alignment data, and polymorphism-derived estimates of per-gene diversity levels. The results demonstrated that these two complementary measures of sequence conservation, one capturing long-term evolutionary constraint and the other assessing current population-level diversity, identify similar biological processes enriched amongst genes at the extremes of their value distributions. Genes expected to be constrained by their key “housekeeping” biological functions exhibit low-diversity and high-constraint levels, while genes with functions implicated in more rapidly evolving processes show high-diversity and low-constraint levels. This was particularly striking for the set of core expression network genes in contrast to the set of genes associated with blood feeding and salivary glands, i.e. gene functional groups defined by their co-expression profiles determined through the meta-analysis of many different experimental gene expression datasets.

The more detailed exploration of relations between conservation and function of immunity genes was performed in the context of a larger study of mosquito immune system evolution that included several other measures of gene evolutionary features (Ruzzante et al., 2022). The results presented here focused on the phastCons constraint metrics and the population-level diversity computed from the MalariaGEN data, with examples presented for antimicrobial peptides, classical recognition proteins, and leucine-rich repeat immune proteins. The results demonstrate that while some functional components of the immune system display low constraint and high diversity levels expected for rapidly and dynamically evolving immunity genes, several other functional categories are instead characterised by genes with much higher levels of sequence conservation. This finding emphasises the main conclusions from Ruzzante *et al.* 2022: "*where and how genes participate in immune responses limit the range of possible evolutionary scenarios they exhibit*" and how "*The test case study system of insect immunity highlights the potential of applying comparative genomics approaches to characterise how functional constraints on different components of biological systems govern their evolutionary trajectories*". Amongst the 18 evolutionary features examined in this paper, evolutionary constraint as measured using phastCons grouped together with gene age, universality, and alignability to form one of the three main axes of evolutionary trajectories. The workflows developed as part of this thesis for building multispecies whole genome alignments (MWGAs), analysing the MWGAs to quantify per-nucleotide or per-gene metrics of conservation, and for visualising the results, will facilitate the future exploration of relationships between gene conservation and gene function for other biological systems and other taxonomic groups.

Chapter 5: Filling technical gaps to implement reproducible, scalable, and portable workflows

Summary

A primary focus of this thesis project was the implementation of reproducible workflows to generate multispecies whole genome alignments (**Chapter 2**) and to compute metrics to estimate conservation of genomic sequences using these alignments (**Chapter 3**). To achieve reproducibility, the workflows are implemented with Snakemake and software dependencies are managed with the Conda package manager. Software used in most steps of the alignment and the analyses workflows was already implemented and ready to be integrated in the workflows; however, the tools required to perform several specific steps did not exist - or not in a state that made them usable in a reproducible workflow. In such cases, rather than writing custom scripts which would be difficult to share and adapt to other scenarios, we chose to develop the missing software ourselves, focusing on open, documented, and extensible implementation in order to make them useful to the community. Furthermore, while some tools were already available as Conda packages, several crucial pieces of software were not available or required fixes to their package to be integrated in our workflows. Over the course of this project, we added multiple packages and fixes to the Bioconda channel, thus contributing to global community efforts to improve portability and reproducibility of scientific software and workflows. Finally, this thesis project - and an increasing number of other projects in our institute - relied heavily on Snakemake. Over time, we implemented a collection of tools and utilities to facilitate implementing complex Snakemake workflows and to execute these workflows on a High Performance Computing (HPC) platform using a SLURM scheduler. In this chapter, we will describe the tools, intermediate workflows, and utilities that were developed and released during this project, providing implementation details when they are relevant, and illustrating our general contribution to the reproducibility of scientific software and workflows.

Implementing the missing blocks of MWGA workflows

An extensible collection of utilities for MWGAs: *mwga-utils*

In part because of how difficult they are to generate, relatively few MWGAs have been computed so far. Several tools and utility scripts have been developed to process alignments in MAF format, including MafFilter (Dutheil et al., 2014), mafTools (Mayakonda et al., 2018), and utilities included in PHAST (Hubisz et al., 2011) and in ROAST (Blanchette et al., 2004). However, there are no community-based, extensive libraries to work with and interface with MAF files, similar to htslib and samtools for reads alignment files or BCFtools for variant annotation files (Bonfield et al., 2021; Danecek et al., 2021). Consequently, several operations on MAF files not covered by existing tools had to be implemented manually; since all these operations rely on reading information from a MAF file, they were implemented in a common software architecture, *mgwa-utils*. This software is implemented in C and C++ and organised around an efficient MAF parser used to implement a collection of utilities; at the moment, four utilities are available: *metrics*, *missing_regions*, *single_cov*, and *stats*, and the design allows to easily implement additional functions.

An efficient MAF parser

The MAF format is defined as a plain text file organised into alignment blocks. It can contain comments line, identified by a starting “#”, and metadata lines, identified by a starting “##”. Each alignment block consists of an Alignment Block line starting with an “a”, which indicates the score of the block, followed by one sequence line starting with an “s” for each sequence in the alignment block. Each sequence line contains six fields separated by a varying number of spaces: 1) assembly and contig from which the sequence originates, 2) start of the sequence on the source contig, 3) end of the sequence on the source contig, 4) strand from which the sequences originates in the source assembly, 5) total size of the contig from which the sequence originates, and 6) the nucleotide sequence in the alignment, including potential insertions. The MAF file is read by the MAF parser function using a C-style buffer; all the information from each line is extracted into a custom MafRecord object, and all the lines in a block are grouped in a MafBlock object. These MafBlock objects are stored in a queue, which enables efficient insertion of new elements and deletion of old elements; batches of alignment blocks are retrieved from the queue and used as the main input for the utilities implemented in *mwga-utils.* There are two advantages to this design: 1) it provides a generic interface with a common input for all analyses using MAF blocks, and 2) it separates MAF reading from the data processing steps, which allows parallelisation of the most compute intensive analyses. Thanks to the efficient C/C++ implementation, the parser is fast and has a low memory footprint. For reference, an alignment file for 19 mosquito species occupying 9.2 Gb of disk space and comprising 89,866,813 total lines for 6,691,196 total alignment blocks is parsed in 166 seconds using a maximum of 9 Mb of memory, without any additional data processing operation, on a standard compute station.

Computing simple sequence conservation metrics on a MAF file

The *metrics* utility from *mwga-utils* implements the computation of simple metrics to estimate conservation of sequence at the nucleotide. At the moment, two metrics are computed for each position in the reference assembly: 1) **alignability**, *i.e.* the number of non-reference assemblies aligned at this position as a frequency between 0 and 1, and 2) **identity**, *i.e.* the proportion of non-reference assemblies with the same nucleotide as the reference, also between 0 and 1. Batches of alignment blocks from the MAF parser are processed in parallel and absolute values for both metrics are stored in a map of contigs and positions. After the entire MAF is processed, all values are divided by the total number of assemblies in the alignment (either given by the user at runtime or estimated from the MAF file) to obtain frequencies. Values for each metric are then exported in wig format, a compact and standard way to represent nucleotide-level values for an entire genome. Using a parsing thread and a single processing thread on the same alignment file described in the previous section, metrics are computed in 12 minutes 43 seconds using a maximum of 14 Gb of memory. Increasing the number of processing threads to eight reduces runtime to 6 minutes 57 seconds and peak memory to 2.3 Gb.

Adding missing reference regions to a MAF file

Genomic regions from the reference assemblies on which no non-reference assembly was aligned are missing from the results of pairwise alignments - only alignment blocks with at least two assemblies are exported. However, most analyses performed on alignment files only consider regions from the reference assembly present in the file. To generate results for the entire reference assembly, including regions that were not aligned to any other assembly, it is thus necessary to add these missing regions to the MAF file. The *missing_regions* utility from *mwga-utils* implements this step, using as input a MAF file and a fasta file for the reference assembly, and outputting a complete MAF file. In practice, the input fasta file for the reference assembly is read to generate 1) a table with the sequence of each contig, and 2) an empty coverage map for each position in each contig in the assembly. Then, the entire input MAF file is parsed and existing blocks are directly written to the MAF output file; genomic positions from the reference assembly are stored in the coverage map. After the entire MAF is processed, the coverage map is used to identify the regions from the reference assemblies that were not present in the original MAF file, and for each of these regions, a single-sequence MAF alignment block is generated and written to the output file. On the 19 mosquitoes alignment described above, the total runtime to add missing reference regions is 3 minutes 15 seconds, with a peak memory usage of 1.8 Gb.

Verifying that all sequences are unique in a MAF file

Genome evolution is a complex process involving a diversity of mechanisms; some of these mechanisms, for instance duplication of genomic regions or loss of genomic content, can impact the number of copies of given genomic sequences and the orthologous relationships of these sequences between species. Although such cases are interesting to understand genomic evolution, their study requires specialised tools to handle copy number variation and non one-to-one orthologous relationships. In practice, most software used to estimate conservation of genomic sequence with MWGAs assumes that sequences from the reference assembly are present only once in the alignment. To achieve this state, the MAF files from pairwise alignments generated in our workflow are processed with the *single_cov2* utility from MULTIZ. To ensure that the multispecies MAF file obtained by combining pairwise alignments also contains only a single copy of each sequence, a *single_cov* utility was implemented in *mwga-utils*. This utility computes the coverage of each genomic position in the reference sequence and outputs a per-contig summary report. On the 19 mosquitoes MAF file described above, the check is completed in 1 minute 47 seconds using a maximum of 1.2 Gb of memory.

Computing simple alignment statistics on a MAF file

Simple statistics on a MAF file can be useful to quickly visualise an alignment or to compare alignment results between different runs, for instance when testing software parameter values. Moreover, the graph visualisation of MWGAs presented in **Chapter 3** requires computation of the number of bases in the alignment for each assembly. This step was implemented in the *stats* utility of *mwga-utils*; in practice, batches of MafBlocks from the queue are processed and the number of bases aligned for each assembly in each block is added to a map of total coverage. After all blocks are processed, a tabulated output file with the name of each assembly and the number of bases in the alignment for the corresponding assembly is generated. The *stats* utility was designed so that other statistics can easily be added in parallel to the one already implemented. On the 19 mosquitoes MAF file, alignment statistics are computed in 1 minute 38 seconds using a maximum of 9 Mb of memory.

Distribution and availability of *mwga-utils*

The *mwga-utils* software is implemented in C++, and all dependencies are provided with the source code at https://github.com/RomainFeron/mwga-utils, so that users can easily build the software for their target platform if needed; the source is distributed under a GPL-3.0 licence. Development milestones are archived as releases, with the current version being 0.1.5; all released versions are available from the Bioconda channel (https://bioconda.github.io/recipes/mwga-utils/README.html).

Leveraging A^{3}Cat to compute phylogenetic trees with Buscophile

Summary of the approach behind Buscophile

One of the inputs required to compute an MWGA is a phylogenetic tree describing the relationship between the species included in the alignment. In order to be able to compute MWGAs for any set of arthropod species, which is one of the goals of this thesis work, we needed a way to generate a phylogenetic tree for the selected set of species. The main challenge in this task was the first step of building phylogenetic trees: acquiring genomic data that can be aligned for all species to include in the tree. To solve this problem, we drew inspiration from *Orthophile*, a workflow developed in our lab which used orthology data from OrthoDB (Kuznetsov et al., 2023) and OMA (Altenhoff et al., 2024) to generate a phylogenetic tree for any set of species included in these databases. We implemented a new version of this workflow, which we called *Buscophile*, using as input the genomic sequences of Universally Single-Copy Orthologues (USCOs) from the results of running BUSCO on all arthropod assemblies for the A^{3}Cat (the Arthropoda assembly assessment catalogue described in **Chapter 1**). This approach relies on the assumption that USCOs are expected to be found in almost all assemblies of sufficient quality to be included in an MWGA, and therefore we can extract a set of alignable genomic sequences for any set of species for which at least one assembly was evaluated in the A^{3}Cat, which by design covers almost all arthropod species sequenced to date.

In practice, *Buscophile* was implemented as a Snakemake workflow which takes as input a set of NCBI TaxIds (Taxon IDs), collects genomic sequences from USCOs shared between the corresponding species from the results of A^{3}Cat, aligns and trims these sequences, and generates a phylogenetic tree from the alignment results. The following sections will provide details on the implementation of this workflow and how we leveraged it to generate trees used to compute MWGAs.

Retrieving USCOs genomic sequences from the A^{3}Cat results

The only input required by *Buscophile* is a list of NCBI TaxIds for the species to include in the tree, provided as a simple text file with one TaxId per line. The first step of the workflow is then to retrieve the latest JSON A^{3}Cat summary from the A^{3}Cat website, using a permanent URL of the *latest* release. Then, for each species, the assembly with the highest Complete BUSCO score as evaluated by A^{3}Cat for a BUSCO dataset specified by the user (default: Arthropoda) is selected as the best assembly for this species and used in the following steps of the workflow. In the same step, a list of IDs for all single-copy complete USCOs found in this assembly is generated from the A^{3}Cat results using a custom script. The result of this step is a text file for each species included in the tree, containing all complete USCO IDs identified in the best assembly for this species. These USCO IDs are collected in the following step to identify USCOs shared by all species included in the tree; USCOs found in a user-defined proportion of these assemblies (default: 95%) are retained and output to a text file with one ID per line. In practice, this threshold can be adjusted based on the number of species included and the quality of the corresponding assemblies. Then, the protein sequences of each retained USCO are extracted from the BUSCO results archive (available from A^{3}Cat) for each assembly into a fasta file containing the sequence of this USCO in each assembly. The final result of the data collection step is a fasta file of protein sequences for each USCO found in a user-defined proportion of the best assemblies for all the species to include in the tree.

Protein sequence alignment and phylogenetic tree reconstruction

Protein sequences obtained in the previous step are aligned separately for each USCO using *muscle* (Edgar, 2022), and the resulting alignment is trimmed with *trimal* (Capella-Gutierrez et al., 2009). The resulting alignments are concatenated into a single alignment file using a custom Python script to produce a single fasta file containing concatenated and aligned sequences of all USCOs in all species. When an USCO is missing from the best assembly for a species, it is input as a gap in the concatenated alignment. The concatenated alignment is used as input to generate a phylogenetic tree; two methods were implemented for this step: *fasttree* (Price et al., 2010) and *iqtree* (Minh et al., 2020). The former is faster but less accurate than the latter, and the user can decide which method is most suitable for their needs. Settings for *fasttree* and *iqtree* can be defined in the config file. The resulting phylogenetic tree is processed to include species names, retrieved from the A^{3}Cat summary file, instead of TaxIds, making the tree more practical to use with recognisable species names.

Practical implementation of the Buscophile workflow

The *Buscophile* workflow was implemented with Snakemake (Köster & Rahmann, 2012), and Conda environments were defined for each step, making the workflow entirely reproducible. A configuration file allows the user to specify parameters for the workflow as well as for the software used to generate phylogenetic trees. The workflow makes use of the resources system described in a later section (Quality of life tools and utilities for Snakemake workflows) to specify runtime and memory usage for each job, thus facilitating running on a computational platform or in a cloud-based environment. The first implementation of the workflow was designed to access A^{3}Cat data directly from the computational platform on which it is run (the UNIL HPC platform), and was thus only suitable for internal use. To make the workflow usable outside our environment, we implemented a second version which accesses data directly from the A^{3}Cat website, thanks to the URL patterns we implemented to download the latest summary table and the BUSCO results archive for any assembly (https://a3cat.unil.ch/downloads.html). The source code for Buscophile is available on Gitlab (https://gitlab.com/evogenlab/buscophile-public).

A Python package to generate GenomeBrowser track hubs: pyGBtracks

One of the final outputs of the *mwgaw-analyses* workflow (described in **Chapter 3**) is a GenomeBrowser track hub to visualise the results across the reference assembly. This hub contains a track for each metric computed by the workflow showing the value of that metric for each position in the reference assembly (alignability, identity, conservation score, and coding potential). Additional tracks display the most conserved elements identified by phastCons as well as putative coding elements suggested by PhyloCSF. The data have to be formatted in a format compatible with GenomeBrowser (Wig, BigWig, Bed, BigBed…) for each track and organised into a track hub folder. Besides the data, this folder contains multiple files describing the hub content and providing metadata about each track for display in GenomeBrowser. In order to automate the task of creating a track hub, we implemented py-GB-tracks, a small Python package that we published on PyPi (https://pypi.org/project/py-GB-tracks). This package implements a *TrackHub* object that creates a hub directory, converts data files if necessary and organises them in the hub directory, and generates metadata files in the hub directory using information provided by the user. The source for py-Gb-tracks is available on Gitlab (https://gitlab.com/evogenlab/mwga/py_gb_tracks), and the package is used as part of the *mwgaw-analyses* workflow.

Contributing to reproducibility of scientific software and workflows

Packaging software for the Bioconda Conda channel

One of the major themes of this thesis work is scientific reproducibility of computational analyses. While reproducibility of the analysis process is achieved using a workflow management system, another challenge is to ensure the exact version of the software used in these analyses can be installed and deployed easily on any supported platform. To this aim, we defined Conda environments for all the tool-requiring steps of our workflows, which are then automatically installed and deployed by Snakemake at runtime. A lot of software used in computational biology is available in the public Bioconda package repository (Grüning et al., 2018), but some tools used in our workflows were not, including our own software like *mwga-utils*. In such cases, rather than including software executables as part of our workflow, we implemented a Conda package for the software and published it in the Bioconda repository. In practice, this process consists in 1) writing a file providing metadata on the software as well as the package itself, and 2) implementing a build system, which can be a simple bash script in the simplest cases or a more complex system in other cases. The package is then extensively tested using Bioconda's build system on both linux and OSX virtual environments and is merged into the Bioconda channel once it successfully builds on both platforms.

Amongst the tools used in our workflows, the following were not available in any Conda package repository: PHAST (Hubisz et al., 2011), MULTIZ (Blanchette et al., 2004), PhyloCSF (Lin et al., 2011), and mwga-utils (**Chapter 5**). We wrote Conda packages for these four pieces of software and published them in the Bioconda package repository; packaging PhyloCSF was particularly challenging as it is written in an old version of OCaml, a seldom-used language with a complex dependencies management system. In addition, we updated the recipe for lastZ (Harris, 2007) to use the correct binary, and we wrote Conda packages for other tools not included in our workflows (for instance *radsex*, *psass*, and *sgtr*, described in **Chapter 6** below).

Quality of life tools and utilities for Snakemake workflows

It should be clear by now that this thesis project made extensive use of the Snakemake workflow management language and engine. Over the course of this project, we accumulated a collection of scripts and utilities implementing commonly-used functions and accessory tools to help develop and run Snakemake workflows. In an effort to make these utilities accessible to the Snakemake user community, we published them in a series of packages that we will describe in the following sections.

A collection of utilities for Snakemake workflows

Our Snakemake workflows make extensive use of Python scripts to process data files. Snakemake provides a way to execute a Python script directly inside a rule; using this method, variables defined in the Snakemake workflow are directly accessible in the Python script via a *snakemake* object. To simplify writing these Python scripts, we implemented commonly-used utility functions in a small Python package, *anguis* (https://github.com/RomainFeron/anguis). One notable feature of *anguis* is a function to redirect Python logs to the log file defined in Snakemake; by default, Python logs were exported to *stdout* and *stderr* and were not collected by Snakemake.

Another feature of Snakemake is the allocation of runtime, memory, and disk usage values for each individual job using the *resources* directive. By default, these values have to be specified either manually for each rule in the workflow, which can be tedious, or once for every rule, which lacks flexibility. To alleviate this issue, we implemented a system which allows to define multiple resource presets for runtime, memory, and disk usage. For instance, memory usage can be defined for each rule using the *large*, *medium*, *small*, or *tiny* preset, and the value for each preset can be set by the user in the workflow's config file based on the requirements of the workflow. Presets are allocated to each rule using a yaml file that is loaded at the beginning of the workflow execution, and values can be overridden for specific rules in the *resources* section of the config file; default values are used when no preset was found for a rule. Our utility functions retrieve preset values from the config file, load resource preset definitions for each rule from the yaml file, and allocate resources for all the rules in the workflow based on this information before executing jobs. In addition, this system automatically increases runtime and memory for a job that is automatically submitted after failure, in case failure was due to a lack of resources.

A Snakemake profile to execute jobs on a SLURM scheduler

To facilitate execution on different systems and using different configurations, Snakemake uses execution profiles, which define a set of Snakemake runtime parameter values. Some of these runtime parameters allow Snakemake to interface with job schedulers commonly found on High-Performance Computing platforms (HPCs), for instance Simple Linux Utility for Resource Management (SLURM), Sun Grid Engine (sge), Oracle Grid Engine, or Portable Batch System (PBS). In particular, runtime parameters allow the user to provide scripts to submit jobs and check job status; therefore, one can implement a Snakemake profile to automate execution on a computational platform. The Snakemake community released such profiles for the most common schedulers, including *slurm*, which is used on the computational platform we accessed for this project (Curnagl). However, this available profile was missing important features required to execute our workflows on Curnagl. In particular, for almost two years, Curnagl was split into two platforms using different configurations but accessed through the same front-end machine and managed by the same scheduler, and thus required careful management of job submission settings to execute workflows automatically. For these reasons, we developed our own Snakemake profile to execute jobs using a SLURM scheduler, implementing missing features we needed to use Curnagl (https://github.com/RomainFeron/snakemake-slurm). Notable features implemented in this profile include automated detection of available partitions every couple days, blacklisting of partitions for job submission, redirection of standard output and error streams to log files defined in the Snakemake workflow, and interfacing with the resources system described in the previous section (*A collection of utilities for Snakemake workflows*) to facilitate resources allocation for individual jobs. This profile was released publicly and maintained over time, and it was used by other members of multiple departments at the university of Lausanne to execute workflows on Curnagl.

Discussion

The work of a bioinformatics or computational biology doctoral thesis will almost always involve the development of technical solutions to achieve various steps of different computational analysis workflows. In many cases however, these technical solutions remain as “quick-fix” answers to solve intermediate hurdles hindering the progress towards completing an analysis in order to achieve the desired results. Such “quick-fixes” may solve short-term problems, but they pose considerable risks for analysis reproducibility and workflow portability. The solutions are often only operational in the hands of the developer and rarely come with documentation, making them difficult to share with others who might be facing similar issues and looking for essentially the same answers to be able to progress with their research. By investing in building sets of technical solutions in a comprehensive manner that makes them sharable and deployable by others, this thesis work contributes to the reproducibility of software and workflows. Specifically, the developed technical solutions for addressing gaps in the provision of reliable methods to perform key steps required to generate MWGAs will greatly enhance the ability of the community to exploit the rapidly growing numbers of available reference genomes for their research. Firstly, they enable researchers to generate and analyse their own MWGAs - a previously extremely challenging operation achievable essentially only by the UCSC team themselves - and secondly, to do this in a reproducible and scalable manner that advances best practices in bioinformatics and computational biology.

Chapter 6: Complementary contributions to genomics research applications

Summary

Over the course of my PhD, I was involved in multiple collaborations which eventually resulted in publications. These projects fall in one of the two following categories: 1) collaborations established during my former position as a Bioinformatician working on teleost sex determination, and 2) projects involving other members of the group or established by my thesis supervisor Robert Waterhouse, focusing on evolutionary genomics of arthropods. In this chapter, I will summarise my contributions to these collaborative projects, focusing first on a major project I lead developing a workflow to study genetic sex determination.

User-friendly tools, visualisations, and workflows to study the genetic mechanisms of sex determination

The central theme of this thesis work is to understand how conservation of sequence across species relates with function and, ultimately, to advance our knowledge on how the information encoded in the genome translates into biological functions. Among all the processes and functions represented across the tree of life, sexual reproduction has received a lot of attention as a central question in the field of evolutionary biology, sometimes being called the “queen of problems” (Pan et al., 2016). Sexual reproduction as a process is extremely conserved across eukaryotes and has many implications in other functions and processes (Beukeboom & Perrin, 2014); in most cases, at least in animals, sexual reproduction involves gonochorism, *i.e.* individuals acquire one of two sexes during development. However, the mechanisms underlying sex determination, the process by which individuals acquire their sex, are highly variable between clades and sometimes even within orders or genera. Sex can be determined non-genetically, using environmental cues like temperature or social cues like relative size in a social group, or genetically, involving a master sex determining gene, *i.e.* the gene at the top of the developmental sex determination cascade, harboured by sex chromosomes. Until recently, most of our knowledge on sex determination came from studies on mammals, avians, and *Drosophila melanogaster*, which share similar genetic sex determination mechanisms with conserved and “degenerated” sex chromosomes accumulating deleterious mutations and eventually losing most of their gene content. This apparent similarity spawned rigid theoretical models of the evolution of sex chromosomes involving their birth after acquiring the master sex determining gene followed by suppression of recombination in a growing region around the sex determining locus. In these models, the region with suppression of recombination extends over time and accumulates deleterious mutations which eventually lead to the loss of most of the sex chromosome, as observed on the mammalian Y chromosome (Graves, 2006). However, as advances in sequencing technologies and in genomics enabled studies of sex determining systems in an increasing number of taxonomically diverse species, it became apparent that these canonical models of sex chromosome evolution described only one scenario among many possible evolutionary trajectories. In particular, studies in Teleost fishes, non-avian reptiles, and in arthropods have uncovered diverse sex determining systems (*e.g.* XX/XO, ZZ/ZO, and polygenic systems), sex chromosome turnover - sometimes even between two sister species, and old sex chromosomes that did not degenerate reviewed in (Bachtrog et al., 2014). These discoveries highlighted the importance of studying the mechanisms of sex determination, sexual reproduction, and in general biological processes in a taxonomically diverse range of species to build evolutionary models that reflect reality. In the case of sex determination, the first step to achieve this goal was to develop genomics tools to identify sex determination systems and characterise sex determining regions in a reproducible way to be able to compare results between different species.

Thanks to personal collaborations that started prior to this PhD thesis, I was involved in multiple projects to study sex determination in Teleost fishes in an evolutionary framework. My role was to develop user-friendly tools and visualisation packages, implement reproducible workflows, and perform genomics analyses to identify the sex determining region and sometimes compare them between species. The development of these tools, workflows, and visualisations, required me to become proficient in the use of software frameworks and best practices to maximise usability and ensure reproducibility. These skills were directly relevant to the work described in **Chapter 1**: producing and maintaining regular updates of the Arthropoda Assembly Assessment Catalogue; **Chapter 2**: computing multispecies whole genome alignments from high-quality genome assemblies; **Chapter 3**: integrating different tools for computing and visualising standardised metrics from multispecies whole genome alignments, and **Chapter 4**: applying the multispecies whole genome alignment workflows to explore functional constraints in mosquito immunity genes.

The first and main output of my complementary contributions to genomics research applications was RADSex, a C++ software to analyse RAD-Sequencing data in order to identify genomic sequences specific to or overrepresented in one sex. While developing RADSex, I focused on user-friendliness and documentation, portability, and performance. In parallel, I implemented an R package, *sgtr*, which provides easy-to-use functions to visualise the results of RADSex as well as other generic functions to plot genomics results, for instance plotting metrics along a genome or along a chromosome. RADSex and *sgtr* were published together with original results on 15 Teleost fish datasets, including five species for which novel sex-biassed sequences were identified. The software and the analyses which I performed are presented in the Associated publication below. Both tools are provided as Conda packages, along with other tools I developed to analyse different types of genomic data, *e.g. PSASS* (https://github.com/SexGenomicsToolkit/PSASS). Whenever relevant and possible, the workflows implementing all the analyses I performed for a publication were made available on GitHub.

Associated publication: RADSex: A computational workflow to study sex determination using restriction site‐associated DNA sequencing data.
Feron et al. Mol Ecol Resour. 2021 Jul;21(5):1715-1731. doi: 10.1111/1755-0998.13360. Epub 2021 Mar 9. PMID: 33590960
GitHub: https://github.com/RomainFeron/paper-sexdetermination-radsex

RADSex, PSASS, *sgtr*, and other tools were used to characterise sex determining regions and sometimes identify candidate master sex determining genes in multiple species of Teleost fishes. In some cases, we were able to apply the reproducible workflows to multiple species within a single order and compare the resulting genomic regions, even retracing the evolutionary history of a master determining gene in the order *Esociformes*. In these projects, which resulted in the associated publications listed below, I was a main contributor to the genomic analyses and I participated in writing and revising of the manuscript.

Associated publication: Characterization of a Y‐specific duplication/insertion of the anti‐Mullerian hormone type II receptor gene based on a chromosome‐scale genome assembly of yellow perch, *Perca flavescens*.
Feron et al. Mol Ecol Resour. 2020 Mar;20(2):531-543. doi: 10.1111/1755-0998.13133. Epub 2020 Jan 27. PMID: 31903688

Associated publication: Identification of the master sex determining gene in Northern pike (*Esox lucius*) reveals restricted sex chromosome differentiation.

Pan et al. PLoS Genet. 2019 Aug 22;15(8):e1008013. doi: 10.1371/journal.pgen.1008013. eCollection 2019 Aug. PMID: 31437150

Associated publication: Independent Origin of XY and ZW Sex Determination Mechanisms in Mosquitofish Sister Species.

Kottler et al. Genetics. 2020 Jan;214(1):193-209. doi: 10.1534/genetics.119.302698. Epub 2019 Nov 8. PMID: 31704715

Associated publication: The rise and fall of the ancient northern pike master sex determining gene.

Pan et al. Elife. 2021 Jan 28;10:e62858. doi: 10.7554/eLife.62858. PMID: 33506762

My experience developing tools and reproducible workflows allowed me to efficiently analyse diverse types of genomic data to identify regions differentiated between males and females. The associated publications listed below represent the outcome of collaborations in which I performed such analyses and contributed to writing the relevant parts of the original manuscript.

Associated publication: The genome of the arapaima (*Arapaima gigas*) provides insights into gigantism, fast growth and chromosomal sex determination system.

Du et al. 2019, Scientific reports 9 (1), 5293. doi: 10.1038/s41598-019-41457-x

Associated publication: The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization.

Du et al. 2020, Nature Ecology & Evolution 4 (6), 841-852. doi: 10.1038/s41559-020-1166-x

Associated publication: Sex chromosome and sex locus characterization in goldfish, Carassius auratus (Linnaeus, 1758).

Wen et al. 2020, BMC Genomics 21 (1), 1-12. doi: 0.1186/s12864-020-06959-3

Associated publication: The bowfin genome illuminates the developmental evolution of ray-finned fishes.

Thompson et al. 2021, Nature Genetics, 1-12. doi: 10.1038/s41588-021-00914-y

Associated publication: A supernumerary “B-sex” chromosome drives male sex determination in the Pachón cavefish, Astyanax mexicanus.

Imarazene et al. 2021, Current Biology 31 (21), 4800-4809. e9. doi: 10.1016/j.cub.2021.08.030
Associated publication: A duplicated copy of id2b is an unusual sex-determining candidate gene on the Y chromosome of arapaima (*Arapaima gigas*).
Adolfi et al. 2021, Scientific Reports 11 (1), 21544. doi: 10.1038/s41598-021-01066-z
Associated publication: Genome biology of the darkedged splitfin, *Girardinichthys multiradiatus*, and the evolution of sex chromosomes and placentation.
Du et al. 2022, Genome Research 32 (3), 583-594. doi: 10.1101/gr.275826.121

Evolutionary genomics in arthropods

In addition to my work to develop tools and workflows and perform analyses to study sex determination, I was involved in several projects investigating evolutionary and / or genomics questions in arthropod clades. In the context of these projects, I computed MWGAs and estimated conservation of sequence, providing data and further downstream analyses results to our collaborators, including other members of the group. I was also involved in quality control for multiple mosquito assemblies, implementing tools to validate scaffolding information from different sources. This work resulted in three publications listed below:

Associated publication: Functional constraints on insect immune system components govern their evolutionary trajectories. *Ruzzante et al. 2022, Molecular biology and evolution 39 (1). doi: 10.1093/molbev/msab352*
This work presents a suite of evolutionary metrics developed to characterise the evolutionary history of genes, and show how these metrics can be used to investigate the relationship between this history and the gene's biological function. In the context of this project, I computed estimates of sequence conservation for all genes in *Anopheles gambiae*, which were included as evolutionary metrics for the study of immune gene families presented in the publication. Computation of these specific sequence conservation metrics was only possible thanks to the workflows I developed to generate and analyse MWGAs.

Associated publication: Anopheles mosquitoes reveal new principles of 3D genome organization in insects. *Lukyanchikova et al 2022, Nature Communications 13 (1). doi: 10.1038/s41467-022-29599-5*
In this study, new assemblies generated for five mosquito species are used to investigate chromatin structural organisation, revealing conserved long-range looping interaction in these

species. I provided insight on the quality of generated assemblies by computing completeness scores using BUSCO.

Associated publication: Evolutionary superscaffolding and chromosome anchoring to improve Anopheles genome assemblies. *Waterhouse et al 2020, BMC Biology 18 (1). doi: 10.1186/s12915-019-0728-3*
This publication uses multiple types of physical chromosome anchoring data to improve the contiguity of draft genome assemblies of 20 mosquito species. I developed the scripts integrating the mapping data to produce scaffolded assemblies from the draft contig-level genomes.

Discussion

The contributions to genomics research applications presented here demonstrate the importance of developing bioinformatics solutions to process and analyse genomics data to advance understanding of biological function and evolution. While they encompass a much broader range of questions than using multispecies whole genome alignments to investigate sequence conservation and function, they share a common theme of tool and workflow development designed to facilitate the exploitation of genomics data. The mainly technical contributions served to advance the research of my colleagues while at the same time delivering methodologies or approaches that can be applied by the wider genomics community. The projects with more intellectual and leadership contributions (RADSex and yellow perch, both published in *Molecular Ecology Resources*) also involved considerable investments in developing tools and analysis workflows, as well as ensuring these were made available for use by the community. The opportunity to participate in diverse research projects from fish to mosquitoes - and even mosquitofish - provided a rich learning environment to develop not only bioinformatics skills but also to engage with key questions in biology and evolution.

Conclusion and perspectives

The work presented in this thesis addresses multiple bioinformatics challenges which had to be overcome in order to explore the relationships between conservation of genomic sequence and biological function of genomic elements in arthropods. The first obstacle encountered was the selection of high quality genome assemblies for several arthropod clades to include in downstream comparative analyses. Indeed, while information on assembled arthropod genomes was generally available, this information was scattered across multiple online resources and databases, or even buried within published articles. Our solution to this problem was the development of a web-based resource presented in **Chapter 1**, the A^{3}Cat, which aggregates available information on existing and upcoming arthropod assemblies and incorporates newly computed estimates of assembly quality generated with a standardised protocol. In addition to this resource, we published the reproducible workflow that we designed to compute the data provided in the A^{3}Cat, as well as additional analyses workflows and packages to generate the website itself. The catalogue is updated monthly to include newly released assemblies, providing users with an up-to-date assessment of the current landscape of arthropod genome assemblies. The A^{3}Cat greatly facilitates the process of selecting assemblies when designing genomics studies and preparing computational analyses; for instance, thanks to the filtering tools implemented in the table, one could easily identify all assemblies released for a species and select the one that best fits their needs based on biological information and assembly quality metrics. The A^{3}Cat also shines when gathering data for the computation of phylogenetic trees by making available and organising the results of executing BUSCO on each assembly in the catalogue; this idea was integrated into a workflow described in **Chapter 5** which enables users to compute a phylogenetic tree for any subset of species included in the catalogue. These features paired with the constant maintenance of the data contributed to the popularity of the A^{3}Cat in the arthropod genomics community: in the year following its release, the resource was accessed by close to 700 unique individuals. Overall, our efforts towards reproducibility of bioinformatics workflows and analyses allowed us to overcome our initial challenge in a way that can benefit the general community, rather than just implementing a simpler private solution for the same purpose. These efforts contributed to improving the findability and accessibility of genomic data for arthropods. It is our hope that in the future the metrics computed for the A^{3}Cat are directly provided by existing resources like NCBI GenBank and the ENA, and that these resources integrate high-level data exploration tools directly in their interface. Until this happens, the A^{3}Cat will uphold this task for arthropods, and the workflows published can be used to generate similar resources for other clades across the tree of life.

The second challenge faced in this thesis was the limited accessibility and usability of computational methods to generate MWGAs, which was likely a major reason preventing their integration into comparative genomics studies. This problem originated from the lack of a publicly available and easily usable workflow for MWGA computation outside of the UCSC GenomeBrowser ecosystem. The same observation was made for downstream analyses using the MWGAs to investigate conservation of sequence and coding potential. This technical challenge spawned the development of two reproducible workflows which are presented in **Chapter 2** and **Chapter 3**. The first workflow, *mwgaw-align*, implements all the steps to compute a MWGA from a set of genome assemblies. The second workflow, *mwgaw-analyses*, provides an automated and user-friendly way to compute estimates of sequence conservation and coding potential using a MWGA, and generates powerful visualisations to explore the results. Both workflows make extensive use of the features provided by workflow the management tools Snakemake and Conda to ensure reproducibility, documentation, scalability, and ease of use; by leveraging the strengths of these tools, we were also able to optimise runtime and memory usage of the most crucial steps, bringing runtime down from potentially weeks to a couple days. Moreover, in addition to addressing the main issue of computing and analysing MWGAs, the workflows were designed to contribute to the ongoing advancement of alignment methodologies thanks to their extensibility and benchmarking capabilities. This is particularly important in light of the resurgence of Cactus to compute MWGAs since the start of this project. While Cactus is progressing rapidly and may become the default approach to compute alignments in the future, the streamlined usability and computational efficiency of *mwgaw-align* will maintain it as a viable alternative promoting healthy competition in alignment method development, thus playing a crucial role in driving innovation and improving the overall quality of methods in this field. In parallel, the *mwgaw-analyses* workflow remains the only automated and reproducible way to efficiently extract insights from a MWGA. The strength of this workflow is its ability to transform the large and obscure MWGA files into a suite of informative outputs: nucleotide-level metrics, gene-level metrics, and visual representations. This feat is achieved thanks to 1) the *mwga-utils* software described in **Chapter 5**, which establishes a framework for computing low-level metrics along MWGAs in MAF file format, 2) the streamlined computation of standard sequence conservation analyses by encapsulating complex tools into a reproducible workflow, and 3) the generation of comprehensive yet concise visualisations of the generated data in the form of genome browser tracks and a web-based table. Similarly to *mwgaw-align*, the extensible design of *mwgaw-analyses* enables easy integration of additional analyses tools, which will contribute to advancing research in this field. Overall, the work presented in these two chapters addressed the need for a reproducible and accessible method to compute MWGAs and execute downstream analyses, while providing support for future advancements in alignment methodologies, thus helping to improve relevance and utility of MWGAs in genomic research. This work represents a significant step towards enabling comprehensive exploration of MWGA data for the larger genomics community beyond the core team at the UCSC, and highlights the importance of reproducibility for both computation and visualisation in genomic analyses.

Each technical chapter in this thesis contributed to building one of the blocks required for large-scale explorations of the relationships between conservation of genomic sequences and the biological function of genomic elements. To validate the viability of this work, we applied them to the case study of conservation of sequence in mosquito genes, comparing the long-term evolutionary constraints estimated from the MWGA with current population polymorphism data available for the reference species *Anopheles gambiae*. This case study, which forms the focus of **Chapter 4**, makes use of all the workflows and visualisations described in the previous chapters and introduces new visualisations that may eventually be implemented in the main workflows. Our analysis of mosquito genes confirms that the genes essential for core biological functions exhibit low diversity and high conservation, while those involved in rapidly evolving processes display the opposite pattern. This was further illustrated by contrasting sets of genes related to core expression networks versus blood feeding and salivary gland co-expressed genes, and through a deeper examination of genes involved in different mosquito immunity functional modules. The insights gained from this preliminary study highlight the interplay between gene function and conservation of genomic sequence. In addition, some of the metrics computed for this work were integrated in a larger study of evolutionary metrics relating to the evolutionary history of genes applied to the same gene set, the results of which underscore the potential of comparative genomics approaches in characterising the evolutionary dynamics of biological systems. In light of this observation, the workflows developed within this thesis for MWGA construction, analysis, and visualisation offer a robust framework for further investigations into the relationship between gene conservation and function across diverse biological systems and taxonomic groups.

As the results of chapter 4 validate our approach at the limited scale of protein-coding genes for a single clade, the next step in extending the work of this thesis project would be scaling up these analyses to potentially hundreds or thousands of genomes generated by sequencing initiatives. Thanks to our development philosophy driven by guidelines for reproducible science, the tools resulting from our work are designed to handle this scaling up. The first task remaining to achieve this goal is combining the outputs of each chapter to complete the following process: 1) identify clades and assemblies suitable for the computation of MWGAs using the A^{3}Cat, 2) gather the necessary data from the A^{3}Cat, including computing a phylogenetic tree with *buscophile*, 3) automatically compute MWGAs and estimate sequence conservation and coding potential over a chosen reference assembly for the selected clades, 4) generate gene-level metrics as well as a list of conserved elements and a list of putative coding elements that can drive future research. The foundation for this complex process was already laid out in a meta-workflow *assemblies2alignments*, but more work is needed before this framework is mature and applicable to all arthropods. Using the resources developed to date, users can already execute each required step to achieve the desired results, but this meta-workflow would provide the framework for integrating all required steps in a common analysis.

The main outputs of this thesis work profoundly change the perspective on the future use and integration of MWGAs in large-scale comparative evolutionary genomics analyses. The methodological advancements in the domain of orthology delineation can serve as an analogy to support this perspective. Many early comparative genomics efforts that relied on cross-species comparisons to identify orthologous genes first developed in-house methods that delivered results but which were rarely, if ever, made public nor usable by others. As methods matured and the numbers of species involved increased, several public databases emerged and provided the community with computed orthology data. Nevertheless, the tools and workflows required to produce these large-scale datasets often remained private to the developers, or if made public they often proved difficult to set up and run, particularly on large datasets. Through community efforts such as the Quest for Orthologs (Altenhoff et al., 2020; Nevers et al., 2022; Sonnhammer et al., 2014), many of the key groups involved concentrated efforts on making publicly available “standalone” software for users to be able to compute their own orthology datasets, to the extent that one of the most popular modern methods is implemented only as a deployable software and without any public database (Emms & Kelly, 2019). These steps towards enabling the wider research community to build their own orthology datasets have democratised the ability to independently generate and use orthology data in their own research projects. The recent use of MWGAs by the Zoonomia consortium has demonstrated the many potential outcomes of exploiting alignment data in comparative genomics analyses to a broad audience of researchers. In this context, the tools and workflows developed through this thesis work present an opportunity for the democratisation of the ability to independently generate and use whole genome alignment data. As such, we believe that MWGAs will increasingly become a feature of comparative genomics research projects that take advantage of the rapidly growing reference genome resources for eukaryotic biodiversity.

Bibliography

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLOS ONE*, *12*(3), e0172792. https://doi.org/10.1371/journal.pone.0172792

Alexa, A., & Rahnenfuhrer, J. (2024). *topGO: Enrichment Analysis for Gene Ontology* (R package version 2.56.0) [Computer software]. Bioconductor. https://doi.org/10.18129/B9.BIOC.TOPGO

Alföldi, J., & Lindblad-Toh, K. (2013). Comparative genomics as a tool to understand evolution and disease. *Genome Research*, *23*(7), 1063–1068. https://doi.org/10.1101/gr.157503.113

Altenhoff, A. M., Garrayo-Ventas, J., Cosentino, S., Emms, D., Glover, N. M., Hernández-Plaza, A., Nevers, Y., Sundesha, V., Szklarczyk, D., Fernández, J. M., Codó, L., for Orthologs Consortium, the Q., Gelpi, J. L., Huerta-Cepas, J., Iwasaki, W., Kelly, S., Lecompte, O., Muffato, M., Martin, M. J., … Dessimoz, C. (2020). The Quest for Orthologs benchmark service and consensus calls in 2020. *Nucleic Acids Research*, *48*(W1), W538–W545. https://doi.org/10.1093/nar/gkaa308

Altenhoff, A. M., Warwick Vesztrocy, A., Bernard, C., Train, C.-M., Nicheperovich, A., Prieto Baños, S., Julca, I., Moi, D., Nevers, Y., Majidian, S., Dessimoz, C., & Glover, N. M. (2024). OMA orthology in 2024: Improved prokaryote coverage, ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA Ecosystem. *Nucleic Acids Research*, *52*(D1), D513–D521. https://doi.org/10.1093/nar/gkad1020

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Angiuoli, S. V., & Salzberg, S. L. (2011). Mugsy: Fast multiple alignment of closely related whole genomes. *Bioinformatics*, *27*(3), 334–342. https://doi.org/10.1093/bioinformatics/btq665

Arcà, B., Lombardo, F., Struchiner, C. J., & Ribeiro, J. M. C. (2017). Anopheline salivary protein genes and gene families: An evolutionary overview after the whole genome sequence of sixteen Anopheles species. *BMC Genomics*, *18*(1), 153. https://doi.org/10.1186/s12864-017-3579-8

Armstrong, J., Hickey, G., Diekhans, M., Fiddes, I. T., Novak, A. M., Deran, A., Fang, Q., Xie, D., Feng, S., Stiller, J., Genereux, D., Johnson, J., Marinescu, V. D., Alföldi, J., Harris, R. S., Lindblad-Toh, K., Haussler, D., Karlsson, E., Jarvis, E. D., … Paten, B. (2020). Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature*, *587*(7833), 246–251. https://doi.org/10.1038/s41586-020-2871-y

Aron, S., Jongeneel, C. V., Chauke, P. A., Chaouch, M., Kumuthini, J., Zass, L., Radouani, F., Kassim, S. K., Fadlelmola, F. M., & Mulder, N. (2021). Ten simple rules for developing bioinformatics capacity at an academic institution. *PLOS Computational Biology*, *17*(12), e1009592. https://doi.org/10.1371/journal.pcbi.1009592

Bachtrog, D., Mank, J. E., Peichel, C. L., Kirkpatrick, M., Otto, S. P., Ashman, T.-L., Hahn, M. W., Kitano, J., Mayrose, I., Ming, R., Perrin, N., Ross, L., Valenzuela, N., Vamosi, J. C., & Consortium, T. T. of S. (2014). Sex Determination: Why So Many Ways of Doing It? *PLOS Biology*, *12*(7), e1001899. https://doi.org/10.1371/journal.pbio.1001899

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. https://doi.org/10.3758/s13428-011-0089-5

Bartholomay, L. C., & Michel, K. (2018). Mosquito immunobiology: The intersection of vector health and vector competence. *Annual Review of Entomology*, *63*(1), 145–167. https://doi.org/10.1146/annurev-ento-010715-023530

Bartholomay, L. C., Waterhouse, R. M., Mayhew, G. F., Campbell, C. L., Michel, K., Zou, Z., Ramirez, J. L., Das, S., Alvarez, K., Arensburger, P., Bryant, B., Chapman, S. B., Dong, Y., Erickson, S. M., Karunaratne, S. H. P. P., Kokoza, V., Kodira, C. D., Pignatelli, P., Shin, S. W., … Muskavitch, M. A. T. (2010). Pathogenomics of Culex quinquefasciatus and meta-analysis of infection responses to diverse pathogens. *Science*, *330*(6000), 88–90. https://doi.org/10.1126/science.1193162

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. https://doi.org/10.1038/483531a

Beukeboom, L. W., & Perrin, N. (2014). *The Evolution of Sex Determination*. Oxford University Press.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., & Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, *14*(4), 708–715. https://doi.org/10.1101/gr.1933104

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L., & Rubin, E. M. (2003). Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome. *Science*, *299*(5611), 1391–1394. https://doi.org/10.1126/science.1081331

Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., & Davies, R. M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. *GigaScience*, *10*(2), giab007. https://doi.org/10.1093/gigascience/giab007

Boomsma, J. J., Brady, S. G., Dunn, R. R., Gadau, J., Heinze, J., Keller, L., Moreau, C. S., Sanders, N. J., Schrader, L., Schultz, T. R., Sundström, L., Ward, P. S., Wcislo, W. T., Zhang, G., & Alliance (GAGA), T. G. A. G. (2017). The Global Ant Genomics Alliance (GAGA). *Myrmecological News*, *25*, 61–66.

Brand, P., Saleh, N., Pan, H., Li, C., Kapheim, K. M., & Ramírez, S. R. (2017). The Nuclear and Mitochondrial Genomes of the Facultatively Eusocial Orchid Bee Euglossa dilemma. *G3 Genes|Genomes|Genetics*, *7*(9), 2891–2898. https://doi.org/10.1534/g3.117.043687

Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Challis, R., Kumar, S., Sotero-Caio, C., Brown, M., & Blaxter, M. (2023). Genomes on a Tree (GoaT): A versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. *Wellcome Open Research*, *8*, 24. https://doi.org/10.12688/wellcomeopenres.18658.1

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (2020). BlobToolKit – Interactive

Quality Assessment of Genome Assemblies. *G3: Genes, Genomes, Genetics*, *10*(4), 1361–1374. https://doi.org/10.1534/g3.119.400908

Childers, A. K., Geib, S. M., Sim, S. B., Poelchau, M. F., Coates, B. S., Simmonds, T. J., Scully, E. D., Smith, T. P. L., Childers, C. P., Corpuz, R. L., Hackett, K., & Scheffler, B. (2021). The USDA-ARS Ag100Pest Initiative: High-Quality Genome Assemblies for Agricultural Pest Arthropod Research. *Insects*, *12*(7), 626. https://doi.org/10.3390/insects12070626

Christmas, M. J., Kaplow, I. M., Genereux, D. P., Dong, M. X., Hughes, G. M., Li, X., Sullivan, P. F., Hindle, A. G., Andrews, G., Armstrong, J. C., Bianchi, M., Breit, A. M., Diekhans, M., Fanter, C., Foley, N. M., Goodman, D. B., Goodman, L., Keough, K. C., Kirilenko, B., … Zhang, X. (2023). Evolutionary constraint and innovation across hundreds of placental mammals. *Science*, *380*(6643), eabn3943. https://doi.org/10.1126/science.abn3943

Christophides, G. K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P. T., Collins, F. H., Danielli, A., Dimopoulos, G., Hetru, C., Hoa, N. T., Hoffmann, J. A., Kanzok, S. M., Letunic, I., Levashina, E. A., Loukeris, T. G., Lycett, G., Meister, S., … Kafatos, F. C. (2002). Immunity-related genes and gene families in Anopheles gambiae. *Science*, *298*(5591), 159–165. https://doi.org/10.1126/science.1077136

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., & Johnston, M. (2003). Finding Functional Features in *Saccharomyces* Genomes by Phylogenetic Footprinting. *Science*, *301*(5629), 71–76. https://doi.org/10.1126/science.1084337

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Bras, Y. L., Lemoine, F., Mareuil, F., Ménager, H., Pradal, C., & Blanchet, C. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, *75*, 284–298. https://doi.org/10.1016/j.future.2017.01.012

Crusoe, M. R., Abeln, S., Iosup, A., Amstutz, P., Chilton, J., Tijanić, N., Ménager, H., Soiland-Reyes, S., Gavrilović, B., Goble, C., & Community, T. C. (2022). Methods included: Standardizing computational reuse and portability with the Common Workflow Language. *Communications of the ACM*, *65*(6), 54–63. https://doi.org/10.1145/3486897

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE*, *5*(6), e11147. https://doi.org/10.1371/journal.pone.0011147

Delcher, A. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, *30*(11), 2478–2483. https://doi.org/10.1093/nar/30.11.2478

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., & Salzberg, S. L. (1999). Alignment of whole genomes. *Nucleic Acids Research*, *27*(11), 2369–2376. https://doi.org/10.1093/nar/27.11.2369

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017).

Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Dohmen, E., Kremer, L. P. M., Bornberg-Bauer, E., & Kemena, C. (2016). DOGMA: Domain-based transcriptome and proteome quality assessment. *Bioinformatics*, *32*(17), 2577–2581. https://doi.org/10.1093/bioinformatics/btw231

Drosophila 12 Genomes Consortium. (2007). Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, *450*(7167), 203–218. https://doi.org/10.1038/nature06341

Dunn, N. A., Unni, D. R., Diesh, C., Munoz-Torres, M., Harris, N. L., Yao, E., Rasche, H., Holmes, I. H., Elsik, C. G., & Lewis, S. E. (2019). Apollo: Democratizing genome annotation. *PLOS Computational Biology*, *15*(2), e1006790. https://doi.org/10.1371/journal.pcbi.1006790

Dutheil, J. Y., Gaillard, S., & Stukenbrock, E. H. (2014). MafFilter: A highly flexible and extensible multiple genome alignment files processor. *BMC Genomics*, *15*(1), 53. https://doi.org/10.1186/1471-2164-15-53

Earl, D., Nguyen, N., Hickey, G., Harris, R. S., Fitzgerald, S., Beal, K., Seledtsov, I., Molodtsov, V., Raney, B. J., Clawson, H., Kim, J., Kemena, C., Chang, J.-M., Erb, I., Poliakov, A., Hou, M., Herrero, J., Kent, W. J., Solovyev, V., … Paten, B. (2014). Alignathon: A competitive assessment of whole-genome alignment methods. *Genome Research*, *24*(12), 2077–2089. https://doi.org/10.1101/gr.174920.114

Edgar, R. C. (2022). Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications*, *13*(1), 6968. https://doi.org/10.1038/s41467-022-34630-w

Elgar, G., & Vavouri, T. (2008). Tuning in to the signals: Noncoding sequence conservation in vertebrate genomes. *Trends in Genetics*, *24*(7), 344–352. https://doi.org/10.1016/j.tig.2008.04.005

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, *306*(5696), 636–640. https://doi.org/10.1126/science.1105136

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Feron, R., & Waterhouse, R. M. (2022a). Assessing species coverage and assembly quality of rapidly accumulating sequenced genomes. *GigaScience*, *11*, giac006. https://doi.org/10.1093/gigascience/giac006

Feron, R., & Waterhouse, R. M. (2022b). Exploring new genomic territories with emerging model insects. *Current Opinion in Insect Science*, *51*, 100902. https://doi.org/10.1016/j.cois.2022.100902

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLOS ONE*, *13*(7), e0200303.

https://doi.org/10.1371/journal.pone.0200303

GIGA Community of Scientists. (2014). The Global Invertebrate Genomics Alliance (GIGA): Developing Community Resources to Study Diverse Invertebrate Genomes. *Journal of Heredity*, *105*(1), 1–18. https://doi.org/10.1093/jhered/est084

Graves, J. A. M. (2006). Sex Chromosome Specialization and Degeneration in Mammals. *Cell*, *124*(5), 901–914. https://doi.org/10.1016/j.cell.2006.02.024

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*(7), Article 7. https://doi.org/10.1038/s41592-018-0046-7

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*. https://etda.libraries.psu.edu/catalog/7971

Harrow, J., Nagy, A., Reymond, A., Alioto, T., Patthy, L., Antonarakis, S. E., & Guigó, R. (2009). Identifying protein-coding genes in genomic sequences. *Genome Biology*, *10*(1), 201. https://doi.org/10.1186/gb-2009-10-1-201

Hecker, N., & Hiller, M. (2020). A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. *GigaScience*, *9*(1), giz159. https://doi.org/10.1093/gigascience/giz159

Herndon, N., Shelton, J., Gerischer, L., Ioannidis, P., Ninova, M., Dönitz, J., Waterhouse, R. M., Liang, C., Damm, C., Siemanowski, J., Kitzmann, P., Ulrich, J., Dippel, S., Oberhofer, G., Hu, Y., Schwirz, J., Schacht, M., Lehmann, S., Montino, A., … Bucher, G. (2020). Enhanced genome assembly and a new official gene set for Tribolium castaneum. *BMC Genomics*, *21*(1), 47. https://doi.org/10.1186/s12864-019-6394-6

Hickey, G., Paten, B., Earl, D., Zerbino, D., & Haussler, D. (2013). HAL: A hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics*, *29*(10), 1341–1342. https://doi.org/10.1093/bioinformatics/btt128

Holm, I., Lavazec, C., Garnier, T., Mitri, C., Riehle, M. M., Bischoff, E., Brito-Fravallo, E., Takashima, E., Thiery, I., Zettor, A., Petres, S., Bourgouin, C., Vernick, K. D., & Eiglmeier, K. (2012). Diverged Alleles of the Anopheles gambiae Leucine-Rich Repeat Gene APL1A Display Distinct Protective Profiles against Plasmodium falciparum. *PLoS ONE*, *7*(12), e52684. https://doi.org/10.1371/journal.pone.0052684

Hoon, S., Ratnapu, K. K., Chia, J., Kumarasamy, B., Juguang, X., Clamp, M., Stabenau, A., Potter, S., Clarke, L., & Stupka, E. (2003). Biopipe: A Flexible Framework for Protocol-Based Bioinformatics Analysis. *Genome Research*, *13*(8), 1904–1915. https://doi.org/10.1101/gr.1363103

Hotaling, S., Sproul, J. S., Heckenhauer, J., Powell, A., Larracuente, A. M., Pauls, S. U., Kelley, J. L., & Frandsen, P. B. (2021). Long-reads are revolutionizing 20 years of insect genome sequencing. *Genome Biology and Evolution*, evab138. https://doi.org/10.1093/gbe/evab138

Hubisz, M. J., Pollard, K. S., & Siepel, A. (2011). PHAST and RPHAST: Phylogenetic analysis with space/time models. *Briefings in Bioinformatics*, *12*(1), 41–51.

https://doi.org/10.1093/bib/bbq072

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*, *33*(6), 1635–1638. https://doi.org/10.1093/molbev/msw046

Hupalo, D., & Kern, A. D. (2013). Conservation and functional element discovery in 20 angiosperm plant genomes. *Molecular Biology and Evolution*, *30*(7), 1729–1744. https://doi.org/10.1093/molbev/mst082

i5K Consortium. (2013). The i5K initiative: Advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *Journal of Heredity*, *104*(5), 595–600. https://doi.org/10.1093/jhered/est050

Jiggins, F. M., & Hurst, G. D. D. (2003). The Evolution of Parasite Recognition Genes in the Innate Immune System: Purifying Selection on Drosophila melanogaster Peptidoglycan Recognition Proteins. *Journal of Molecular Evolution*, *57*(5), 598–605. https://doi.org/10.1007/s00239-003-2506-6

Jungreis, I., Chan, C. S., Waterhouse, R. M., Fields, G., Lin, M. F., & Kellis, M. (2016). Evolutionary dynamics of abundant stop codon readthrough. *Molecular Biology and Evolution*, *33*(12), 3108–3132. https://doi.org/10.1093/molbev/msw189

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., & Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, *423*(6937), 241–254. https://doi.org/10.1038/nature01644

Kent, W. J. (2002). BLAT —The BLAST -Like Alignment Tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Kern, F., Fehlmann, T., & Keller, A. (2020). On the lifetime of bioinformatics web services. *Nucleic Acids Research*, *48*(22), 12523–12533. https://doi.org/10.1093/nar/gkaa1125

Khan, Y. A., Jungreis, I., Wright, J. C., Mudge, J. M., Choudhary, J. S., Firth, A. E., & Kellis, M. (2020). Evidence for a novel overlapping coding sequence in POLG initiated at a CUG start codon. *BMC Genetics*, *21*(1), 25. https://doi.org/10.1186/s12863-020-0828-7

Kheradpour, P., Stark, A., Roy, S., & Kellis, M. (2007). Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Research*, *17*(12), 1919–1931. https://doi.org/10.1101/gr.7090407

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493. https://doi.org/10.1101/gr.113985.110

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kuang, J., Buchon, N., Michel, K., & Scoglio, C. (2022). A global Anopheles gambiae gene co-expression network constructed from hundreds of experimental conditions with missing values. *BMC Bioinformatics*, *23*(1), 170. https://doi.org/10.1186/s12859-022-04697-9

Kuznetsov, D., Tegenfeldt, F., Manni, M., Seppey, M., Berkeley, M., Kriventseva, E. V., & Zdobnov, E. M. (2023). OrthoDB v11: Annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Research*, *51*(D1), D445–D451. https://doi.org/10.1093/nar/gkac998

Kwon, H., Mohammed, M., Franzén, O., Ankarklev, J., & Smith, R. C. (2021). Single-cell analysis of mosquito hemocytes identifies signatures of immune cell subtypes and cell differentiation. *eLife*, *10*, e66192. https://doi.org/10.7554/eLife.66192

Lazzaro, B. P., Zasloff, M., & Rolff, J. (2020). Antimicrobial peptides: Application informed by evolution. *Science*, *368*(6490), eaau5480. https://doi.org/10.1126/science.aau5480

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., Hoad, G., Jang, M., Pakseresht, N., Plaister, S., Radhakrishnan, R., Reddy, K., Sobhany, S., Ten Hoopen, P., Vaughan, R., … Cochrane, G. (2011). The European Nucleotide Archive. *Nucleic Acids Research*, *39*(suppl_1), D28–D31. https://doi.org/10.1093/nar/gkq967

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, *115*(17), 4325–4333. https://doi.org/10.1073/pnas.1720115115

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lian, Y., Yan, C., Xu, H., Yang, J., Yu, Y., Zhou, J., Shi, Y., Ren, J., Ji, G., & Wang, K. (2018). A Novel lncRNA, LINC00460, Affects Cell Proliferation and Apoptosis by Regulating KLF2 and CUL4A Expression in Colorectal Cancer. *Molecular Therapy - Nucleic Acids*, *12*, 684–697. https://doi.org/10.1016/j.omtn.2018.06.012

Liang, P., Saqib, H. S. A., Zhang, X., Zhang, L., & Tang, H. (2018). Single-base resolution map of evolutionary constraints and annotation of conserved elements across major grass genomes. *Genome Biology and Evolution*. https://doi.org/10.1093/gbe/evy006

Lin, M. F., Carlson, J. W., Crosby, M. A., Matthews, B. B., Yu, C., Park, S., Wan, K. H., Schroeder, A. J., Gramates, L. S., St. Pierre, S. E., Roark, M., Wiley, K. L., Kulathinal, R. J., Zhang, P., Myrick, K. V., Antone, J. V., Celniker, S. E., Gelbart, W. M., & Kellis, M. (2007). Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Research*, *17*(12), 1823–1836. https://doi.org/10.1101/gr.6679507

Lin, M. F., Jungreis, I., & Kellis, M. (2011). PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, *27*(13), i275–i282. https://doi.org/10.1093/bioinformatics/btr209

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., … Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, *478*(7370), 476–482. https://doi.org/10.1038/nature10530

Liu, G., Zhang, H., Zhao, C., & Zhang, H. (2020). Evolutionary History of the Toll-Like Receptor Gene Family across Vertebrates. *Genome Biology and Evolution*, *12*(1), 3615–3634. https://doi.org/10.1093/gbe/evz266

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Margulies, E. H., Blanchette, M., Program, N. C. S., Haussler, D., & Green, E. D. (2003).

Identification and Characterization of Multi-Species Conserved Sequences. *Genome Research*, *13*(12), 2507–2518. https://doi.org/10.1101/gr.1602203

Margulies, E. H., Cooper, G. M., Asimenos, G., Thomas, D. J., Dewey, C. N., Siepel, A., Birney, E., Keefe, D., Schwartz, A. S., Hou, M., Taylor, J., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Brown, J. B., Bickel, P., … Sidow, A. (2007). Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Research*, *17*(6), 760–774. https://doi.org/10.1101/gr.6034307

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics*, *7*, 29–59. https://doi.org/10.1146/annurev.genom.7.080505.115623

Mathe, C. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, *30*(19), 4103–4117. https://doi.org/10.1093/nar/gkf543

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., & Koeffler, H. P. (2018). Maftools: Efficient and comprehensive analysis of somatic variants in cancer. *Genome Research*, *28*(11), 1747–1756. https://doi.org/10.1101/gr.239244.118

Merkel, D. (2014). *Docker: Lightweight linux containers for consistent development and deployment*. *2014*(239).

MetaEuk—Sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics | Microbiome. (n.d.). Retrieved May 6, 2024, from https://link.springer.com/article/10.1186/s40168-020-00808-x

Miller, W., Rosenbloom, K., Hardison, R. C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D. C., Baertsch, R., Blankenberg, D., Pond, S. L. K., Nekrutenko, A., Giardine, B., Harris, R. S., Tyekucheva, S., Diekhans, M., Pringle, T. H., Murphy, W. J., Lesk, A., … Kent, W. J. (2007). 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Research*, *17*(12), 1797–1808. https://doi.org/10.1101/gr.6761107

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Minkin, I., & Medvedev, P. (2020). Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *Nature Communications*, *11*(1), 6327. https://doi.org/10.1038/s41467-020-19777-8

Mudge, J. M., Jungreis, I., Hunt, T., Gonzalez, J. M., Wright, J. C., Kay, M., Davidson, C., Fitzgerald, S., Seal, R., Tweedie, S., He, L., Waterhouse, R. M., Li, Y., Bruford, E., Choudhary, J. S., Frankish, A., & Kellis, M. (2019). Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci. *Genome Research*, *29*(12), 2073–2087. https://doi.org/10.1101/gr.246462.118

Myllymäki, H., & Rämet, M. (2014). JAK/STAT pathway in Drosophila immunity. *Scandinavian Journal of Immunology*, *79*(6), 377–385. https://doi.org/10.1111/sji.12170

Nassar, L. R., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, B. T., Lee, C. M., Muthuraman, P., Nguy, B., Pereira, T., Nejad, P., Perez, G., Raney, B. J., Schmelter, D., Speir, M. L., … Kent, W. J.

(2023). The UCSC Genome Browser database: 2023 update. *Nucleic Acids Research*, *51*(D1), D1188–D1195. https://doi.org/10.1093/nar/gkac1072

Neafsey, D. E., Waterhouse, R. M., Abai, M. R., Aganezov, S. S., Alekseyev, M. A., Allen, J. E., Amon, J., Arcà, B., Arensburger, P., Artemov, G., Assour, L. A., Basseri, H., Berlin, A., Birren, B. W., Blandin, S. A., Brockman, A. I., Burkot, T. R., Burt, A., Chan, C. S., … Besansky, N. J. (2015). Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. *Science*, *347*(6217), 1258522. https://doi.org/10.1126/science.1258522

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nevers, Y., Jones, T. E. M., Jyothi, D., Yates, B., Ferret, M., Portell-Silva, L., Codo, L., Cosentino, S., Marcet-Houben, M., Vlasova, A., Poidevin, L., Kress, A., Hickman, M., Persson, E., Piližota, I., Guijarro-Clarke, C., the OpenEBench team the Quest for Orthologs Consortium, Altenhoff, A., Bruford, E. A., … Altenhoff, A. (2022). The Quest for Orthologs orthology benchmark service in 2022. *Nucleic Acids Research*, *50*(W1), W623–W632. https://doi.org/10.1093/nar/gkac330

Nola, R., & Sankey, H. (2014). *Theories of Scientific Method: An Introduction*. Routledge. https://doi.org/10.4324/9781315711959

Nuijten, M. B., Hartgerink, C. H. J., Van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Obbard, D. J., Welch, J. J., Kim, K.-W., & Jiggins, F. M. (2009). Quantifying Adaptive Evolution in the Drosophila Immune System. *PLOS Genetics*, *5*(10), e1000698. https://doi.org/10.1371/journal.pgen.1000698

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A., & Li, P. (2004). Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, *20*(17), 3045–3054. https://doi.org/10.1093/bioinformatics/bth361

Ovcharenko, I., Loots, G. G., Giardine, B. M., Hou, M., Ma, J., Hardison, R. C., Stubbs, L., & Miller, W. (2005). Mulan: Multiple-sequence local alignment and visualization for studying function and evolution. *Genome Research*, *15*(1), 184–194. https://doi.org/10.1101/gr.3007205

Palatini, U., Masri, R. A., Cosme, L. V., Koren, S., Thibaud-Nissen, F., Biedler, J. K., Krsticevic, F., Johnston, J. S., Halbach, R., Crawford, J. E., Antoshechkin, I., Failloux, A.-B., Pischedda, E., Marconcini, M., Ghurye, J., Rhie, A., Sharma, A., Karagodin, D. A., Jenrette, J., … Bonizzoni, M. (2020). Improved reference genome of the arboviral vector Aedes albopictus. *Genome Biology*, *21*(1), 215. https://doi.org/10.1186/s13059-020-02141-w

Pan, Q., Anderson, J., Bertho, S., Herpin, A., Wilson, C., Postlethwait, J. H., Schartl, M., & Guiguen, Y. (2016). Vertebrate sex-determining genes play musical chairs. *Comptes Rendus. Biologies*, *339*(7–8), 258–262. https://doi.org/10.1016/j.crvi.2016.05.010

Pang, K. C., Frith, M. C., & Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. *Trends in Genetics*, *22*(1), 1–5.

https://doi.org/10.1016/j.tig.2005.10.003

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*(9), 1061–1067. https://doi.org/10.1093/bioinformatics/btm071

Paten, B., Earl, D., Nguyen, N., Diekhans, M., Zerbino, D., & Haussler, D. (2011). Cactus: Algorithms for genome multiple sequence alignment. *Genome Research*, *21*(9), 1512–1528. https://doi.org/10.1101/gr.123356.111

Pockrandt, C., Steinegger, M., & Salzberg, S. L. (2022). PhyloCSF++: A fast and user-friendly implementation of PhyloCSF with annotation tools. *Bioinformatics*, *38*(5), 1440–1442. https://doi.org/10.1093/bioinformatics/btab756

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, *20*(1), 110–121. https://doi.org/10.1101/gr.097857.109

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rao, X.-J., Zhan, M.-Y., Pan, Y.-M., Liu, S., Yang, P.-J., Yang, L.-L., & Yu, X.-Q. (2018). Immune functions of insect βGRPs and their potential application. *Developmental & Comparative Immunology*, *83*, 80–88. https://doi.org/10.1016/j.dci.2017.12.007

Reijnders, M. J. M. F., & Waterhouse, R. M. (2021). Summary Visualizations of Gene Ontology Terms With GO-Figure! *Frontiers in Bioinformatics*, *1*, 638255. https://doi.org/10.3389/fbinf.2021.638255

Rottschaefer, S. M., Riehle, M. M., Coulibaly, B., Sacko, M., Niaré, O., Morlais, I., Traoré, S. F., Vernick, K. D., & Lazzaro, B. P. (2011). Exceptional Diversity, Maintenance of Polymorphism, and Recent Directional Selection on the APL1 Malaria Resistance Genes of Anopheles gambiae. *PLoS Biology*, *9*(3), e1000600. https://doi.org/10.1371/journal.pbio.1000600

Roux, J., Privman, E., Moretti, S., Daub, J. T., Robinson-Rechavi, M., & Keller, L. (2014). Patterns of Positive Selection in Seven Ant Genomes. *Molecular Biology and Evolution*, *31*(7), 1661–1685. https://doi.org/10.1093/molbev/msu141

Ruzzante, L., Feron, R., Reijnders, M. J. M. F., Thiébaut, A., & Waterhouse, R. M. (2022). Functional Constraints on Insect Immune System Components Govern Their Evolutionary Trajectories. *Molecular Biology and Evolution*, *39*(1), msab352. https://doi.org/10.1093/molbev/msab352

Ruzzante, L., Reijnders, M. J. M. F., & Waterhouse, R. M. (2019). Of genes and genomes: Mosquito evolution and diversity. *Trends in Parasitology*, *35*(1), 32–51. https://doi.org/10.1016/j.pt.2018.10.003

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2019). GenBank. *Nucleic Acids Research*, *47*(D1), D94–D99. https://doi.org/10.1093/nar/gky989

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., & Miller, W. (2003). Human–Mouse Alignments with BLASTZ. *Genome Research*, *13*(1), 103–107. https://doi.org/10.1101/gr.809403

Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., & Miller, W. (2000). PipMaker—A Web Server for Aligning Two Genomic DNA Sequences. *Genome Research*, *10*(4), 577–586. https://doi.org/10.1101/gr.10.4.577

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., & Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Siepel, A., & Haussler, D. (2005). Phylogenetic Hidden Markov Models. In *Statistical Methods in Molecular Evolution* (pp. 325–351). Springer-Verlag. https://doi.org/10.1007/0-387-27733-1_12

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*(1), 195–197. https://doi.org/10.1016/0022-2836(81)90087-5

Sonnhammer, E. L. L., Gabaldón, T., Sousa da Silva, A. W., Martin, M., Robinson-Rechavi, M., Boeckmann, B., Thomas, P. D., & Dessimoz, C. (2014). Big data and other challenges in the quest for orthologs. *Bioinformatics*, *30*(21), 2993–2998. https://doi.org/10.1093/bioinformatics/btu492

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics*, *13*(9), Article 9. https://doi.org/10.1038/nrg3207

Stallman, R., McGrath, R., & Smith, P. D. (2004). *GNU Make: A program for directing recompliation ; GNU make version 3.81*. Free Software Foundation.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(suppl_2), W435–W439. https://doi.org/10.1093/nar/gkl200

Stark, A., Kheradpour, P., Parts, L., Brennecke, J., Hodges, E., Hannon, G. J., & Kellis, M. (2007). Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Research*, *17*(12), 1865–1879. https://doi.org/10.1101/gr.6593807

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J., Hodges, E., Hinrichs, A. S., Caspi, A., Paten, B., Park, S.-W., Han, M. V., Maeder, M. L., … Kellis, M. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, *450*(7167), 219–232. https://doi.org/10.1038/nature06340

The Anopheles gambiae 1000 Genomes Consortium. (2021). *Ag1000G phase 3 SNP data release. MalariaGEN.* [dataset].

https://www.malariagen.net/data_package/ag1000g-phase3-snp/.

The Darwin Tree of Life Project Consortium. (2022). Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences*, *119*(4), e2115642118. https://doi.org/10.1073/pnas.2115642118

The ENCODE Project Consortium. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*(7146), 799–816. https://doi.org/10.1038/nature05874

The Galaxy Community, Afgan, E., Nekrutenko, A., Grüning, B. A., Blankenberg, D., Goecks, J., Schatz, M. C., Ostrovsky, A. E., Mahmoud, A., Lonie, A. J., Syme, A., Fouilloux, A., Bretaudeau, A., Nekrutenko, A., Kumar, A., Eschenlauer, A. C., DeSanto, A. D., Guerler, A., Serrano-Solano, B., … Briggs, P. J. (2022). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, *50*(W1), W345–W351. https://doi.org/10.1093/nar/gkac247

Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., Blanchette, M., Siepel, A. C., Thomas, P. J., McDowell, J. C., Maskeri, B., Hansen, N. F., Schwartz, M. S., Weber, R. J., Kent, W. J., Karolchik, D., Bruen, T. C., Bevan, R., Cutler, D. J., … Green, E. D. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, *424*(6950), 788–793. https://doi.org/10.1038/nature01858

Upton, L. M., Povelones, M., & Christophides, G. K. (2015). Anopheles gambiae blood feeding initiates an anticipatory defense response to Plasmodium berghei. *Journal of Innate Immunity*, *7*(1), 74–86. https://doi.org/10.1159/000365331

Ureta-Vidal, A., Ettwiller, L., & Birney, E. (2003). Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nature Reviews Genetics*, *4*(4), 251–262. https://doi.org/10.1038/nrg1043

Valanne, S., Wang, J.-H., & Rämet, M. (2011). The Drosophila Toll signaling pathway. *The Journal of Immunology*, *186*(2), 649–656. https://doi.org/10.4049/jimmunol.1002302

Wang, L., & Jiang, T. (1994). On the Complexity of Multiple Sequence Alignment. *Journal of Computational Biology*, *1*(4), 337–348. https://doi.org/10.1089/cmb.1994.1.337

Wang, Q., Ren, M., Liu, X., Xia, H., & Chen, K. (2019). Peptidoglycan recognition proteins in insect immunity. *Molecular Immunology*, *106*, 69–76. https://doi.org/10.1016/j.molimm.2018.12.021

Wang, X., Wang, J., Jin, D., Guo, H., Lee, T.-H., Liu, T., & Paterson, A. H. (2015). Genome Alignment Spanning Major Poaceae Lineages Reveals Heterogeneous Evolutionary Rates and Alters Inferred Dates for Key Evolutionary Events. *Molecular Plant*, *8*(6), 885–898. https://doi.org/10.1016/j.molp.2015.04.004

Waterhouse, R. M., Kriventseva, E. V., Meister, S., Xi, Z., Alvarez, K. S., Bartholomay, L. C., Barillas-Mury, C., Bian, G., Blandin, S., Christensen, B. M., Dong, Y., Jiang, H., Kanost, M. R., Koutsos, A. C., Levashina, E. A., Li, J., Ligoxygakis, P., MacCallum, R. M., Mayhew, G. F., … Christophides, G. K. (2007). Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science*, *316*(5832), 1738–1743. https://doi.org/10.1126/science.1139862

Waterhouse, R. M., Povelones, M., & Christophides, G. K. (2010). Sequence-structure-function relations of the mosquito leucine-rich repeat immune proteins. *BMC Genomics*, *11*(1),

531. https://doi.org/10.1186/1471-2164-11-531

Waterhouse, R. M., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2019). Using BUSCO to assess insect genomic resources. In S. J. Brown & M. E. Pfrender (Eds.), *Insect Genomics* (Vol. 1858, pp. 59–74). Springer New York. https://doi.org/10.1007/978-1-4939-8775-7_6

Waterhouse, R. M., Zdobnov, E. M., Tegenfeldt, F., Li, J., & Kriventseva, E. V. (2011). OrthoDB: The hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Research*, *39*(suppl_1), D283–D288. https://doi.org/10.1093/nar/gkq930

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wilusz, J. E., Sunwoo, H., & Spector, D. L. (2009). Long noncoding RNAs: Functional surprises from the RNA world. *Genes & Development*, *23*(13), 1494–1504. https://doi.org/10.1101/gad.1800909

Woodard, S. H., Fischman, B. J., Venkat, A., Hudson, M. E., Varala, K., Cameron, S. A., Clark, A. G., & Robinson, G. E. (2011). Genes involved in convergent evolution of eusociality in bees. *Proceedings of the National Academy of Sciences*, *108*(18), 7472–7477. https://doi.org/10.1073/pnas.1103457108

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y. J. K., Cooke, J. E., & Elgar, G. (2004). Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biology*, *3*(1), e7. https://doi.org/10.1371/journal.pbio.0030007

Wright, M. W., & Bruford, E. A. (2011). Naming “junk”: Human non-protein coding RNA (ncRNA) gene nomenclature. *Human Genomics*, *5*(2), 90. https://doi.org/10.1186/1479-7364-5-2-90

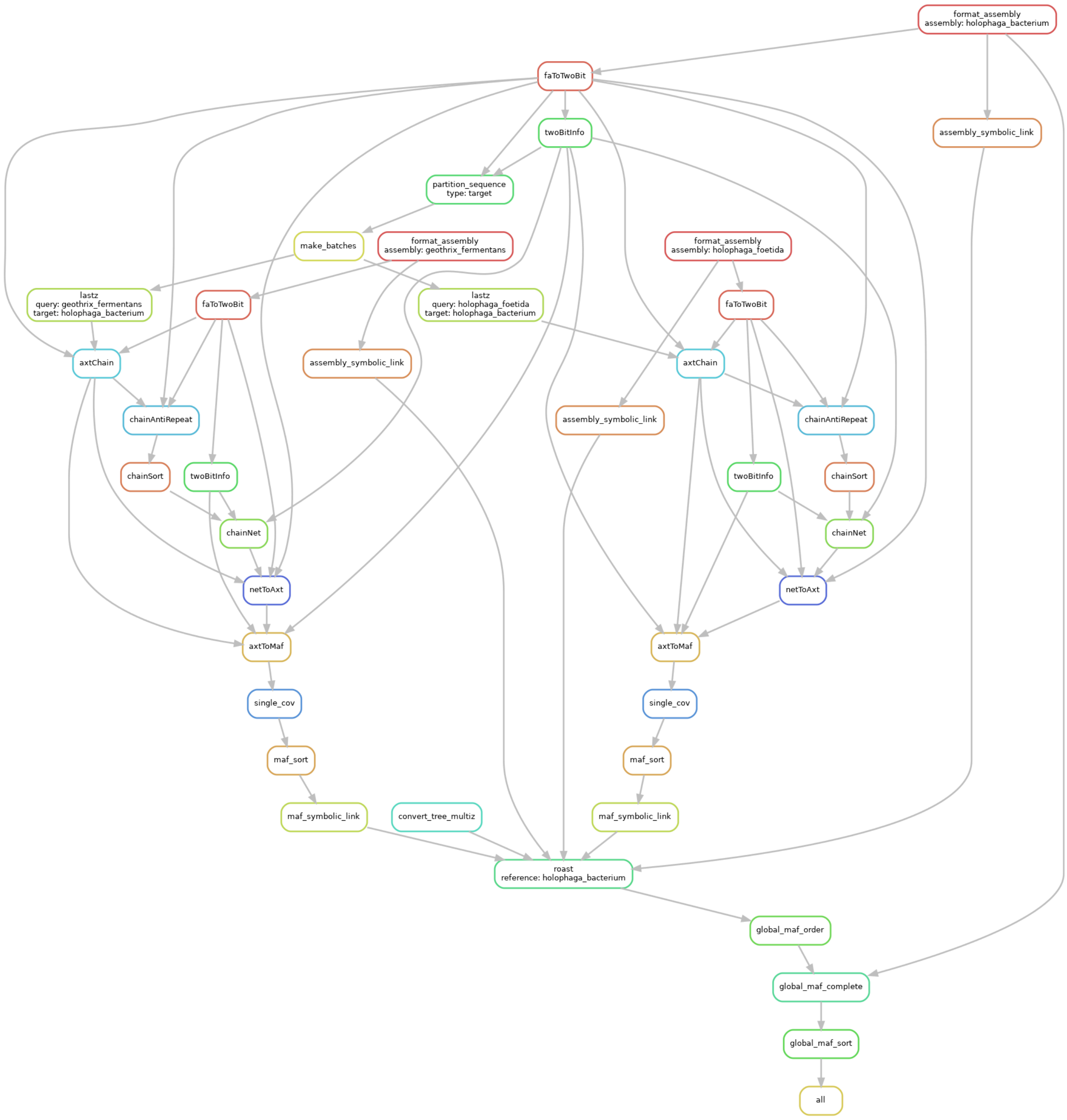
Wyman, S. K., Jansen, R. K., & Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, *20*(17), 3252–3255. https://doi.org/10.1093/bioinformatics/bth352

Ye, Y., Zhang, H., Li, D., Zhuo, J., Shen, Y., Hu, Q., & Zhang, C. (2021). Chromosome‐level assembly of the brown planthopper genome with a characterized Y chromosome. *Molecular Ecology Resources*, *21*(4), 1287–1298. https://doi.org/10.1111/1755-0998.13328

Zhang, L., Liu, B., Zheng, W., Liu, C., Zhang, D., Zhao, S., Li, Z., Xu, P., Wilson, K., Withers, A., Jones, C. M., Smith, J. A., Chipabika, G., Kachigamba, D. L., Nam, K., d’Alençon, E., Liu, B., Liang, X., Jin, M., … Xiao, Y. (2020). Genetic structure and insecticide resistance characteristics of fall armyworm populations invading China. *Molecular Ecology Resources*, *20*(6), 1682–1696. https://doi.org/10.1111/1755-0998.13219

Zoonomia Consortium. (2020). A comparative genomics multitool for scientific discovery and conservation. *Nature*, *587*(7833), 240–245. https://doi.org/10.1038/s41586-020-2876-6

Appendix 3



Appendix 3 Figure 1: complete Directed Acyclic Graph (DAG) representation of the *mwgaw-align* workflow for three bacterial assemblies, automatically generated with Snakemake. The graph shows parallel processing of assemblies but does not include parallel processing of batches, which use the advanced *dynamic checkpoints* Snakemake feature which cannot be represented in the DAG.