Journal of Apicultural Research

ISSN: 0021-8839 (Print) 2078-6913 (Online) Journal homepage: www.tandfonline.com/journals/tjar20

Standard methods and good practices in *Apis* honey bee omics research

Maeva A. Techer, Priyadarshini Chakrabarti, Lílian Caesar, Sonia E. Eynard, M. Catherine Farrell, Leonard J. Foster, June Gorrochategui-Ortega, Dora Henriques, Hongmei Li-Byarlay, Jeffrey T. Morré, Irene L. G. Newton, Melanie Parejo, M. Alice Pinto, Alain Vignal, Iratxe Zarraonaindia & Alison McAfee

To cite this article: Maeva A. Techer, Priyadarshini Chakrabarti, Lílian Caesar, Sonia E. Eynard, M. Catherine Farrell, Leonard J. Foster, June Gorrochategui-Ortega, Dora Henriques, Hongmei Li-Byarlay, Jeffrey T. Morré, Irene L. G. Newton, Melanie Parejo, M. Alice Pinto, Alain Vignal, Iratxe Zarraonaindia & Alison McAfee (2025) Standard methods and good practices in *Apis* honey bee omics research, Journal of Apicultural Research, 64:2, 307-402, DOI: 10.1080/00218839.2025.2455852

To link to this article: https://doi.org/10.1080/00218839.2025.2455852

JOURNAL OF APICULTURAL RESEARCH
2025, VOL. 64, NO. 2, 307–402
https://doi.org/10.1080/00218839.2025.2455852

REVIEW ARTICLE

Standard methods and good practices in *Apis* honey bee omics research

Maeva A. Techer[a,b,c], Priyadarshini Chakrabarti[d,e*], Lílian Caesar[f], Sonia E. Eynard[g,h], M. Catherine Farrell[i], Leonard J. Foster[j], June Gorrochategui-Ortega[k], Dora Henriques[l,m], Hongmei Li-Byarlay[i,n], Jeffrey T. Morré[o], Irene L. G. Newton[f], Melanie Parejo[k], M. Alice Pinto[l,m], Alain Vignal[g], Iratxe Zarraonaindia[k,p] and Alison McAfee[j,q]

[a]Okinawa Institute of Science and Technology, Okinawa, Japan; [b]Department of Entomology, Texas A&M University, College Station, TX, USA; [c]Behavioral Plasticity Research Institute, NSF-BII, Houston, TX, USA; [d]Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Mississippi State, MS, USA; [e]Department of Horticulture, Oregon State University, Corvallis, OR, USA; [f]Department of Biology, Indiana University, Bloomington, IN, USA; [g]GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, France; [h]Labogena DNA, Jouy-en-Josas, France; [i]Agricultural Research and Development Program, Central State University, Wilberforce, OH, USA; [j]Department of Biochemistry and Molecular Biology, Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada; [k]Department of Genetics, Physical Anthropology and Animal Physiology, Applied Genomics and Bioinformatics, University of the Basque Country (UPV/EHU), Leioa, Spain; [l]Centro de Investigação de Montanha (CIMO), LA SusTEC, Instituto Politécnico de Bragança, Campus de Santa Apolónia, Bragança, Portugal; [m]Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, Bragança, Portugal; [n]Department of Agricultural and Life Sciences, Central State University, Wilberforce, OH, USA; [o]Department of Chemistry, Oregon State University, Corvallis, OR, USA; [p]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain; [q]Department of Applied Ecology, North Carolina State University, Raleigh, NC, USA

ABSTRACT

In the past decades, COLOSS members have joined forces multiple times to develop and condense standard methods related to research on honey bees, their pests, pathogens, and colony products. This led to the publication of four open-access *BEEBOOK* volumes that have been utilized by researchers worldwide. Among the chapters, "Standard methods for molecular research in *Apis mellifera*," written by Evans and collaborators in 2013, has been a cornerstone for the standardization of honey bee molecular studies. However, since sequencing technologies and analyzing algorithms have made tremendous progress, many described methods require updating. In parallel, other *Apis* species' genomes have now been sequenced, thus opening new research avenues in a comparative framework. In this chapter, we add to the methods previously covered by Evans et al. in 2013 and provide updated methodology where necessary, including worked examples and bioinformatic analysis pipelines. We also cover topics which were not previously covered in depth, such as sequencing ancient samples, population genomics, proteomics, and sampling honey bee colony products for microbiome studies, among others. Our hope is for this to become a lasting resource for honey bee scientists as the field continues to advance.

MÉTODOS ESTÁNDAR Y BUENAS PRÁCTICAS EN LA INVESTIGACIÓN ÓMICA DE LAS ABEJAS DEL GÉNERO *APIS*

En las últimas décadas, los miembros de COLOSS han unido sus fuerzas en múltiples ocasiones para desarrollar y condensar métodos estándar relacionados con la investigación de las abejas melíferas, sus plagas, patógenos y productos apícolas. Esto ha dado lugar a la publicación de cuatro volúmenes *BEEBOOK* de acceso libre que han sido utilizados por investigadores de todo el mundo. Entre los capítulos, «Métodos estándar para la investigación molecular en *Apis* mellifera», escrito por Evans y colaboradores en 2013, ha sido una piedra angular para la estandarización de los estudios moleculares de las abejas melíferas. Sin embargo, dado que las tecnologías de secuenciación y los algoritmos de análisis han avanzado enormemente, muchos de los métodos descritos requieren una actualización. Paralelamente, se han secuenciado los genomas de otras especies de *Apis*, lo que abre nuevas vías de investigación en un marco comparativo. En este capítulo, revisamos los métodos previamente cubiertos por Evans et al. en 2013 y proporcionamos metodología actualizada cuando es necesario, incluyendo ejemplos trabajados y pipelines de análisis bioinformático. También cubrimos temas que antes no se trataban en profundidad, como la secuenciación de muestras antiguas, la genómica de poblaciones, la proteómica y el muestreo de productos de colonias de abejas melíferas para estudios del microbioma, entre otros. El marco de este trabajo también está disponible en forma de wiki que puede actualizarse en tiempo

ARTICLE HISTORY
Received 4 July 2023
Accepted 2 September 2024

KEYWORDS
COLOSS; *BEEBOOK*; *Apis*; honey bee; omics; standard methods; protocol

CONTACT Maeva A. Techer maeva.techer@ag.tamu.edu; Alison McAfee alison.n.mcafee@gmail.com
*Current affiliation: Washington State University affiliation Department of Entomology, Washington State University, Pullman, WA 99164, USA.
Please refer to this paper as: Techer, M.A., Chakrabarti, P., Caesar, L., Eynard, S.E., Farrell, M.C., Foster, L.J., Gorrochategui-Ortega, J., Henriques, D., Li-Byarlay, H., Morré, J.T., Newton, I.L.G., Parejo, M., Pinto, M.A., Vignal, A., Zarraonaindia, I., McAfee, A. (2025) Standard methods and good practices in *Apis* honey bee 'omics research. In *P. Chantawannakul, J.D. Evans, P. Neumann, N. L. Carreck, J. D. Ellis & V. Dietemann (Eds.), The COLOSS BEEBOOK, Volume IV: Standard methods for Apis cerana research and* Apis *'omics. Journal of Apicultural Research*, 64(2). https://doi.org/10.1080/00218839.2025.2455852

Supplemental data for this article can be accessed online at https://doi.org/10.1080/00218839.2025.2455852.

real a medida que se produzcan futuros avances tecnológicos. Esperamos que se convierta en un recurso duradero para los científicos de las abejas melíferas, a medida que el campo siga evolucionando.

蜜蜂组学研究的标准方法和良⊠做法

在过去的几十年中，COLOSS成员多次联合开发和凝练与蜜蜂 、蜜蜂病害和蜂产 品等研究相关的标 准方法。出版了 四本可免费浏览的*BEEBOOK*， 供世界各地的研究人员使 用。其中埃文斯及合作者于 2013年撰写的一个章节"意 大利蜜蜂 分子研究的标准方 法"一直是蜜蜂分子研究标准化的基石。然而， 由于测序技术和分析算法取得了巨大的进步， 许多描述的方法需要更新。与 此同时， 蜜蜂属其它物种的基因组也已完成测序， 从而在比较框架下开辟了新的研究 途径。在本章中，我们将重温埃文斯等人在 2013 年 介绍的方法，并在必要时提供更新的方法， 包括工作实例和生物信息学分析管 道。我们也涵盖了 以前没 有深入探讨的主题，如古代样本测序、种群基因组学、蛋白质组学、微生物组研究中的蜂产品采样等。本文的框架也可以通过维基站点获得， 并随 未来技术的进步 而实时更新。我们希望随着该领域的不断发展，将其做成蜜蜂科学家的一个持久性资源。

Table of contents

1. The 'omics revolution in *Apis*: more data than meets the eye

While digging into our evolutionary history through archeology, we have found that humans have interacted with bees for at least 40,000 years, revealing a profound and intricate connection (d'Errico et al., 2012). In the native range of the honey bee *Apis mellifera*, beeswax was utilized in pottery during the Neolithic agricultural revolution (Roffet-Salque et al., 2015). Given this ancient relationship and associated benefits, it came as no surprise that the western honey bee genome was among the first insects to be sequenced in 2006 (Honey bee Genome Sequencing Consortium, 2006; Toth & Zayed, 2021). This breakthrough paved the way for multiple research avenues and applications for studying *A. mellifera* evolution, biology, behavior, genetics, conservation, and health, which have been extensively reviewed by Toth and Zayed (2021) and Grozinger and Zayed (2020). The *Apis* genus (Apidae, Hymenoptera), however, encompasses at least ten other species showing an incredible diversity of adaptations in Asia and Oceania (Panziera et al., 2022; Radloff et al., 2010; Randall Hepburn & Radloff, 2011).

Toth and Zayed (2021) have thoroughly summarized the explosive growth of honey bee genetics and genomics studies since 2006. Early on, short DNA and RNA sequences obtained via Sanger sequencing (∼100–1500 bp) regularly enriched the *A. mellifera* GenBank database. Now, progressively high-throughput and next-generation sequencing (NGS) technologies produce millions of sequences per individual and must be compiled into "digestible" Sequence Read Archive (SRA) format. This massive volume of genetic data can be overwhelming and is likely underutilized. For example, our search in the SRA Run Selector of NCBI (accessed on 23 January 2023), yielded 15,458 hits strictly associated with "(*Apis*[Organism]) AND '*Apis mellifera*' [orgn:_txid7460]" (Figure 1(A)). This creates opportunities for processing and comparing large data sets of *A. mellifera* but requires a standardized baseline for future comparative questions and analyses.

The actual SRA size generated from *A. mellifera* genomes and metagenomes is surely underestimated here, due to inconsistencies in the metadata

reports. Evans et al. (2013) anticipated the comparison hardships and discordances that could arise from such a global burst of data and responded with the first set of standard methods and molecular toolkits for *A. mellifera*. The timely publication of the resulting *BEEBOOK* chapter (Evans et al., 2013) preceded the release of 99.2% ($n = 8,185$) and 94.1% ($n = 4,715$) of the total genomic and transcriptomic SRA available at the time of our survey, respectively. This emphasizes the need of standardizing not only the upstream data generation processes (e.g., sampling, wet lab processing and sequencing) but also downstream processes (such as sharing and facilitating open access distribution online).

If *A. mellifera* remains the winning and most curated species within the *Apis* genus, its sister species *A. cerana* displays an early similar growth trajectory (Figure 1(B)). Since 2011, multiple and improved versions of five *Apis* species genomes have been sequenced (Figure 2), offering new opportunities for the comparison of their biology and genomes. The development of new *BEEBOOK* chapters dedicated to *A. cerana* and the further recognition of its unique evolutive history and biological traits are likely to drive a burst in data generation, analysis, and sharing. Fortunately, many of the molecular standard methods are applicable beyond *A. mellifera* and, in some cases, transferrable

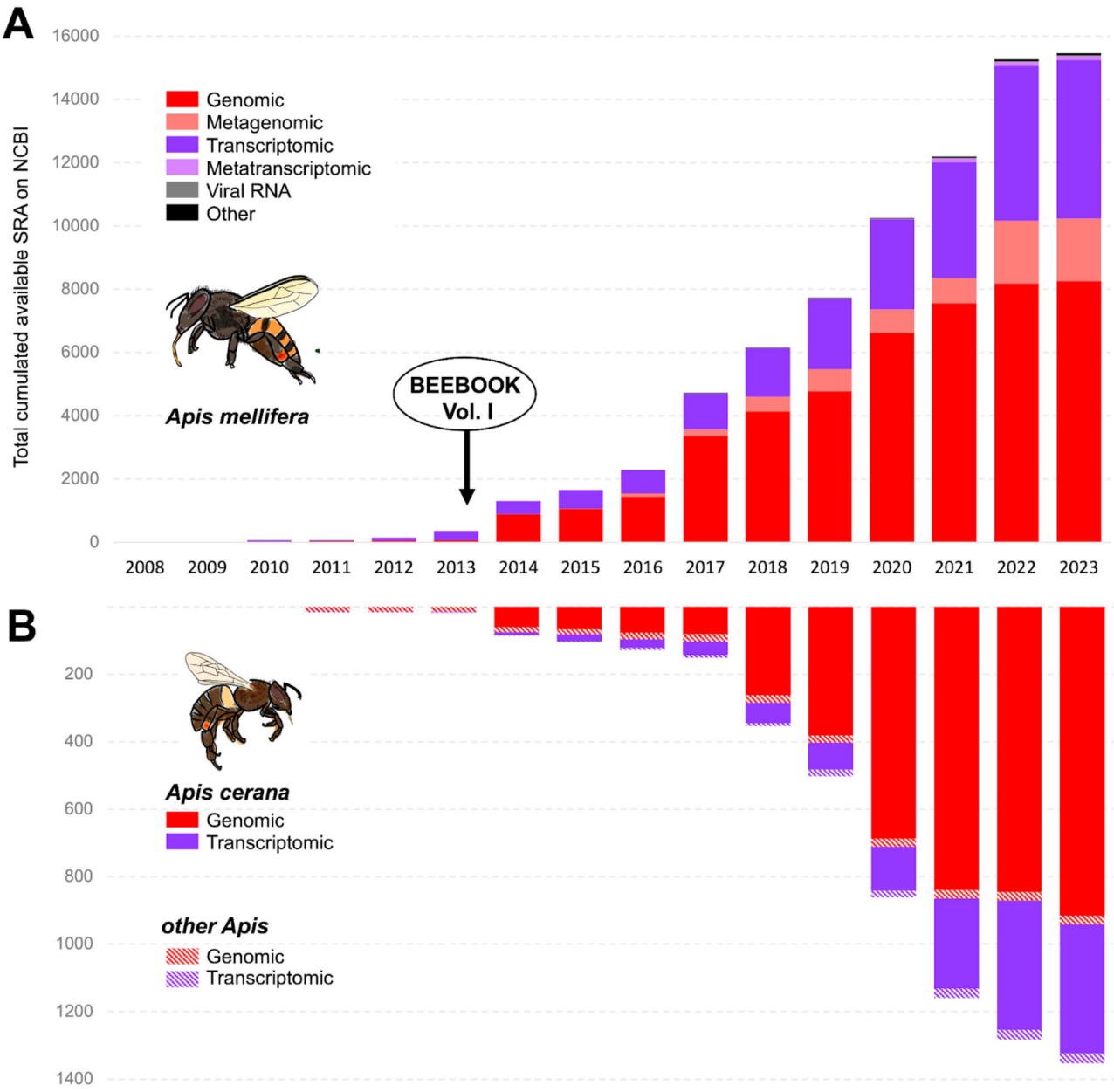

Figure 1. The cumulated short read archive on NCBI for (A) "*Apis mellifera*" or (B) "*Apis*" *stricto* sensu reflects the burst in genetic and genomic resources since the release of the western honey bee genome in 2006.

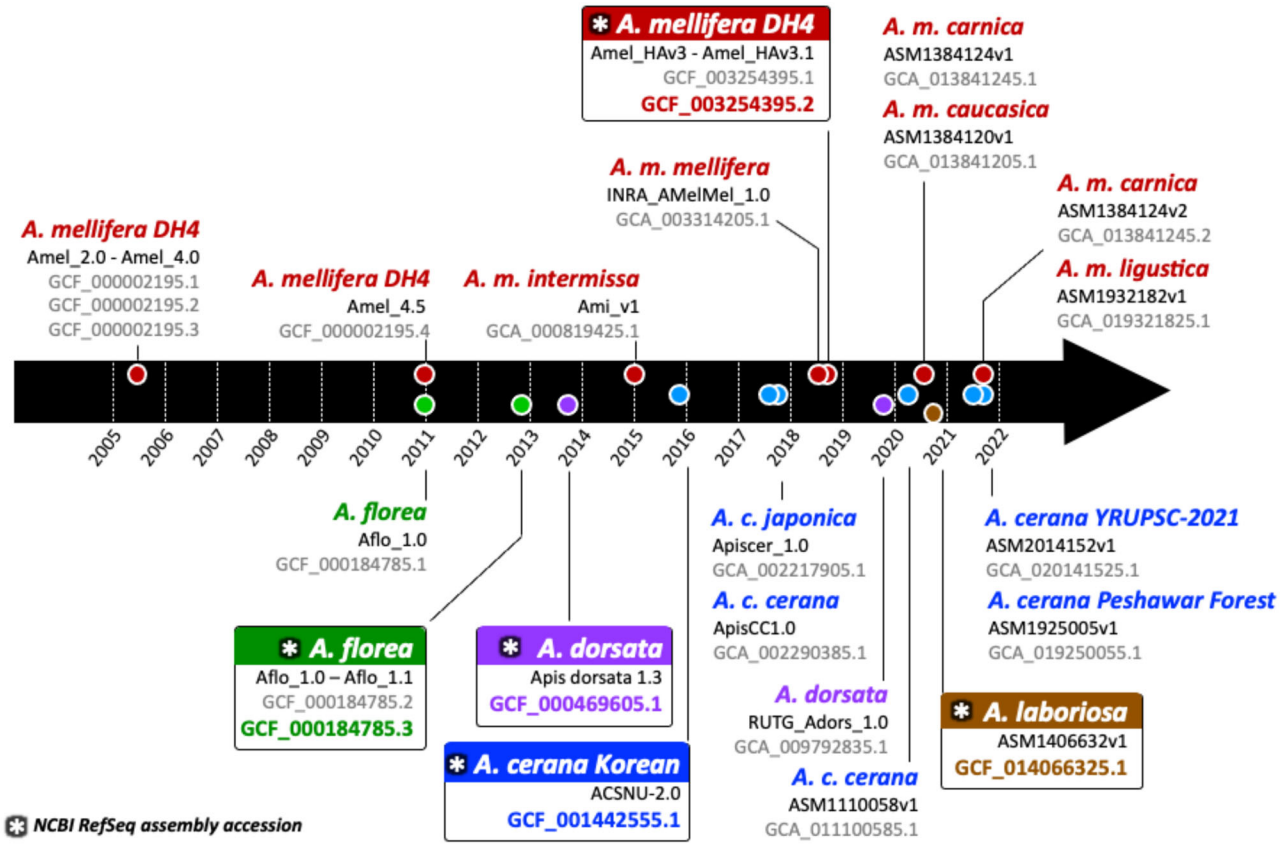

Figure 2. Five reference genomes of *Apis* honey bee species have been assembled and improved since the first genome release. The current timeline of the assemblies publicly released (e.g., ACSNU-2.0 [GFC_001442555.1]), as well as the representative genome (RefSeq) for each species and subspecies, was built using the latest NCBI update.

to other arthropods (Childers et al., 2021; Lawniczak et al., 2022).

In parallel, the emergence of new omics techniques and steady progress in wet lab and bioinformatics techniques urged a revisit and expansion of the Evans et al. (2013) chapter. The data generated by epigenomics, proteomics, and metabolomics are also expanding resources, and it becomes crucial to understand how to analyze, use and share them efficiently. Ultimately, *Apis* research will move toward multi-omics integration, and it will become crucial to efficiently utilize unique or rare samples for multiple layers of data generation and analysis. We encourage niche-specialized *Apis* researchers to familiarize themselves with the outputs and capabilities of each omic method to fully leverage the wealth of accessible data. Leveraging our experience with the transient nature of omics technologies and bioinformatics pipelines, we have enhanced this *BEEBOOK* chapter by also making it available as an interactive wiki (available at https://maevatecher.github.io/standard-methods-apis-omics/; doi.org/10.5281/zenodo.14697986) (Deligkaris, 2022). We compiled the most up-to-date methods while also describing the applications and limitations. Finally, we offer recommendations for the standardization of data sharing in view of the omics future in *Apis*. Our hope is that this chapter will become a lasting resource as the technologies continue to advance.

2. Sample management

Sample management for honey bee samples is essentially the same as described for *A. mellifera* in *BEEBOOK* volume II (Evans et al., 2013). In cases where samples may need to be collected and handled differently before processing (e.g., museum samples), we indicate deviations at the beginning of each protocol.

3. Genome sequencing

3.1. Introduction

As for other species, the sequencing of *Apis* DNA has many applications that can be divided into three categories: de novo sequencing, resequencing, and transcript sequencing. While the last category technically relies on sequencing cDNA reads, its use is intended to inform the structure and expression of genes in honey bee genome, a topic that is covered in Section 6.

De novo sequencing of *Apis* species started in 2006 but is punctually used to generate improved

reference genomes and to represent new subspecies (Figure 2) (Toth & Zayed, 2021). To achieve successful de novo sequencing, the utilization of the most advanced technology available is necessary. The longest possible reads should be produced, possibly from one single sample, and these will be assembled into contigs based on partial sequence overlap (Figure 3(A)). In turn, these contigs will be assembled using other mapping methods such as optical maps or Hi-C, to reconstruct larger fragments (scaffolds), aiming at a chromosome-level assembly.

Further analyses, such as population genomics or RNA-seq, may then use the reference genome produced by de novo sequencing. This is done by aligning reads produced with a high-throughput method, such as Illumina short-read parallel sequencing, to the reference. Using a single reference genome for a community of users allows the comparison of results by having an unique coordinate system. For instance, population genomics studies on whole genomes are conducted by a re-sequencing approach in which the reads from one sample are compared to a reference by alignment (Figure 3(B)). Re-sequencing with short reads will detect small differences such as single nucleotide polymorphisms (SNPs) or insertion-deletion mutations (indels), whereas a long-read approach will highlight larger structural variants. Alignment of sequence reads to a reference is also used for a variety of other analyses, such as the detection of DNA methylation (bisulfite sequencing) (Lyko et al., 2010), identifying regulatory regions linked to histone modifications (ChIP-seq) (Nakato & Sakata, 2021) the analysis of the 3D conformation of chromosomes in the interphase nucleus (Hi-C) (Hoencamp et al., 2021; van Berkum et al., 2010), and many other applications.

Different sequencing techniques exist, and the choice amongst the three main categories (Sanger sequencing, short read parallel sequencing, and long-read sequencing) will depend on the desired goal. Due to the broad applications of sequencing and the constant progress made by technology, we will only cover the most common ones as quick guides toward informed choices here. For reviews on the three main sequencing technologies, see (Heather & Chain, 2016; Shendure et al., 2017).

3.2. Genome sequencing technologies

3.2.1. Sanger sequencing

Until the mid-2000s, DNA sequencing primarily relied on the Sanger technique, which was invented in the 1970s and partially automated in the late 1980s with the introduction of sequencing machines. These first-generation sequencers represented great progress at the time and were based on the size fractionation of DNA fragments by electrophoresis and laser detection of the four possible bases by using four fluorochromes. However, for each 500–1000 bp read

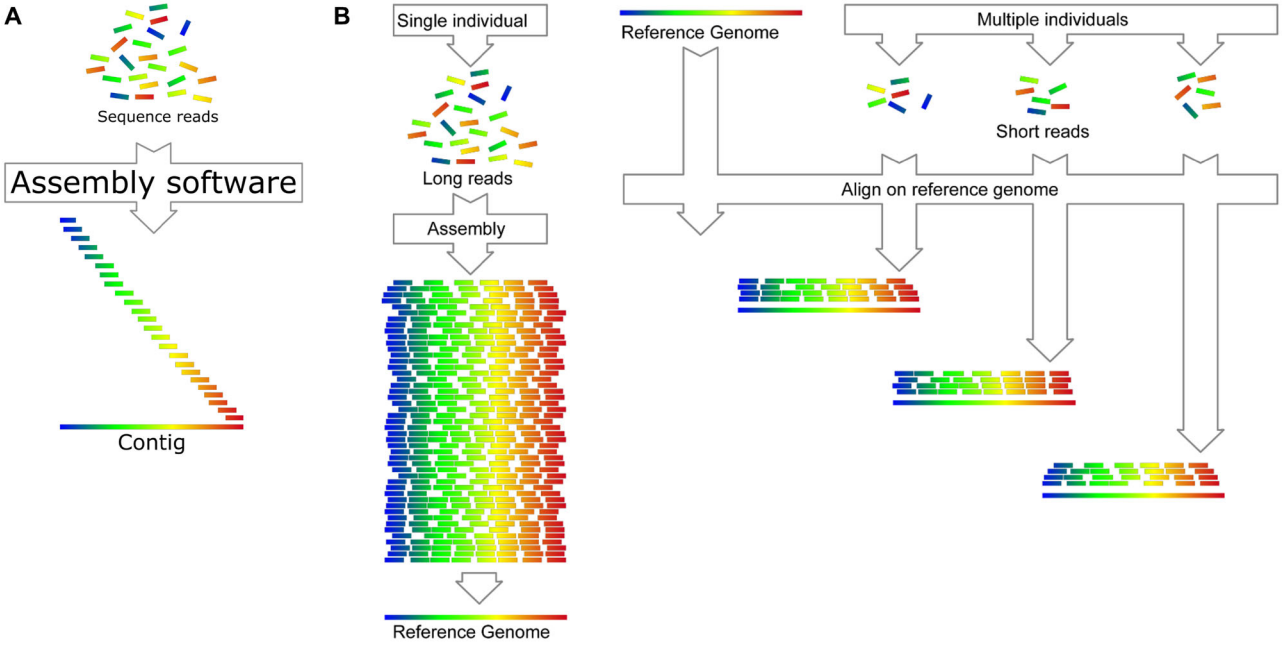

Figure 3. Genome sequencing. (A) Sequence reads are assembled into contigs by partial sequence overlap. (B) Alignment of reads to a reference genome. These can be whole-genome sequencing, RNA-seq or other. Ideally, for a given *Apis* species, only one reference genome should be used by the community, allowing for a unique and common coordinate system for comparing results. Sequencing genomes usually refers to this method of looking for differences between the samples under study and the reference. The consequence is that all results are reference-biased. For instance, a gene absent in the reference genome cannot be analyzed in the samples, even when reads align to it.

produced, separate sequencing reactions, based on the copy of a template DNA, had to be performed. This technique has become obsolete in favor of next-generation genome sequencing, but is still used for sequencing polymerase chain reaction (PCR) amplicons when targeting specific regions of one honey bee genome.

3.2.2. Next generation sequencing

The next major genome sequencing breakthrough was the advent of next-generation sequencing (NGS) techniques in the mid-2000s. Although these were proposed at first by three companies (Roche, Applied Biosystems, and Solexa/Illumina), today, the dominant platform is Illumina. The breakthrough came from the fact that the sequencing reactions were no longer performed individually, but simultaneously on a surface, or flow cell. This allows millions of DNA fragments to be amplified in parallel, with fluorescently labeled nucleotides added and detected sequentially. Parallel sequencing has a very high throughput and can currently produce up to billions of reads per run. However, these reads are short (150–250 bp, depending on the technology used), which can be a major limitation, especially for de novo sequencing. This inconvenience is partially overcome by the fact that two 150 bp reads (read pairs) can be produced from both ends of each DNA fragment. Before the advent of long-read sequencing, read pairs distant up to 10 kb could be produced (mate-pairs) to help in sequence assembly and scaffolding. Parallel sequencing is used, for instance, in population genomics or for generating a very high density of unbiased markers in a genome-wide association study.

3.2.3. Long-read sequencing

Long-read sequencing, pioneered by Pacific Biosciences (PacBio) and Oxford Nanopore, is the newest sequencing approach. These innovative technologies can produce reads longer than 10 kb, but until recently, at the cost of a high sequence error rate. As of the time of writing, both parallel and long-read sequencing are the technologies of choice and are used either independently or in combination. Long-read sequencing is often used for producing new genome assemblies and for the detection of structural variants (SVs). For a detailed discussion on using long-read sequencing for transcriptomics, please refer to Section 6.2.2.

Today, sequencing is done via dedicated core facilities or private companies. Users submit their samples or DNA, and in return, are supplied with the sequencing files along with quality assessment of the sequenced data. Most of the work, then, consists of analyzing the data to extract biological meaning. However, depending on the biological question and perhaps also on budget considerations, the sequencing strategy (e.g., read depth, platform) will have to be defined in advance, and at least some basic knowledge of the advantages and limits of the current technologies are required. Sequencing platforms often provide tools to guide the new user (e.g., coverage calculator, sample pooling normalization calculator) but we do recommend consulting sequencing specialists as a very first step.

3.3. The reference genome

Genomic analyses, such as genome-wide association studies (GWAS), population genomics, and transcriptomics are virtually impossible to perform in the absence of a reference genome for the species studied. For instance, SNPs are detected by aligning sequence reads from samples to the reference and looking for the differences, transcriptome analyses align sequence reads from RNA samples to the reference genome, and transcript levels are determined by counting reads mapped onto the different annotated genes. Moreover, the accuracy of such analyses depends highly on the quality of the reference genome. A typical example in honey bees is the gene number estimation that went from a first Official Gene Set (OGSv1.0) of 10,157 protein-coding genes in version 2 of the assembly (The Honeybee Genome Sequencing Consortium, 2006) to a much larger 15,314 protein-coding OGSv3.2 gene set detected in Amel_4.5 (Elsik et al., 2014). Among many reasons for this difference are the progress in DNA sequencing techniques and assembly algorithms. However, Amel_4.5 remained a very fragmented assembly. More recently, the utilization of long-read sequencing technologies has led to the development of an updated and highly contiguous genome assembly (HAv3.1) which has corrected many errors in chromosome segment ordering (Wallberg et al., 2019). Such a gapless assembly is also essential for accurate gene annotation (Denton et al., 2014). Using a single reference genome for all subsequent analyses will allow having a consistent coordinate system, which is indispensable for comparing results.

3.3.1. Assembling the reference genome

The reference genome should be as perfect as possible, and maximum effort must be made to use state-of-the-art technologies and bioinformatics strategies. As of the time of writing, the current honey bee genome build, HAv3.1, was produced using PacBio long-read sequencing, and the reads were assembled into contigs with `FALCON` (version 0.5.0). The contigs were then merged using additional information, mainly genetic maps, BioNano optical

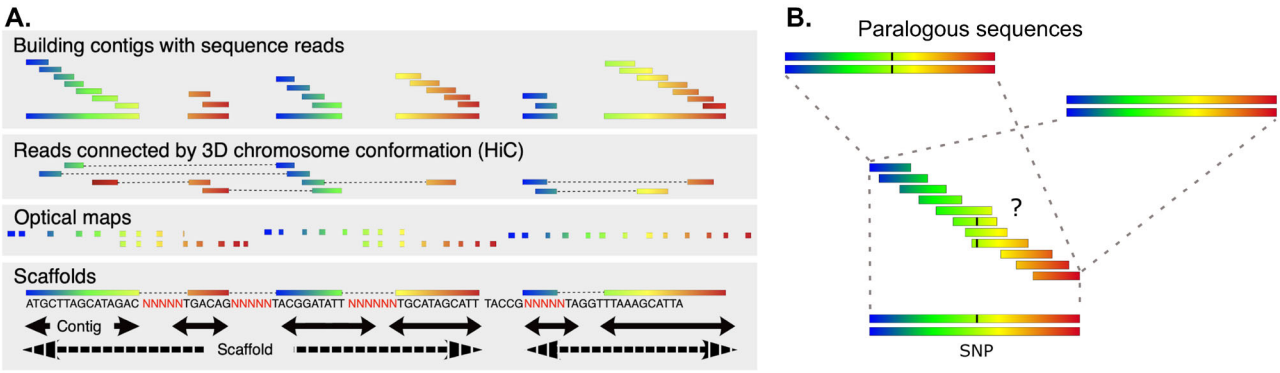

Figure 4. Genome assembly. (A) Contigs are first built by sequence overlap and connected together into scaffolds with linked reads, usually using HiC and/or optical maps. (B) De novo assembly decisions can be influenced by polymorphism present in the sample used as reference. When reconstructing a haploid consensus sequence from a diploid individual, a certain proportion of mismatches must be allowed, to take polymorphisms into account. However, the presence of paralogous sequences in the genome will complicate the decision process. Sequencing a haploid drone solves this problem.

maps, and Hi-C chromatin interaction data (Figure 4(A)) (Wallberg et al., 2019).

3.3.1.1. Hi-C chromosome conformation capture. The Hi-C chromosome conformation capture method was developed for analyzing the 3D organization of the genome, including possible interactions between distant loci, either on the same chromosome or different chromosomes (van Berkum et al., 2010). However, most interactions involve relatively close loci, following the compaction of the DNA in the chromatin and topologically associated domains (TADs). Therefore, Hi-C is also used to detect read pairs that will map at distances in the order of 10–100 kb to help assemble contigs together in whole genome de novo sequencing. Hi-C method is particularly valuable in assembling regions of the genome that are challenging using conventional sequencing methods alone (e.g., due to inversions, large chromosomes, gaps and repetitive elements). To reveal the chromatin looping, Hi-C was utilized to compare the 3D genome structures of queen and worker larvae (Zhang, He, et al., 2023). Hi-C and PacBio technologies are also used to generate the chromosome-scale assembly of the *Apis cerana* and *A. mellifera* genomes (Cao et al., 2021; Wallberg et al., 2019; Wang et al., 2020).

3.3.1.2. DNA source selection. For technical reasons at first, and now for continuity reasons, all reference genome assemblies for the honey bee were done using inbred queens or drones from the DH4 strain; Bee Weaver Apiaries, Inc. (Elsik et al., 2014; The Honeybee Genome Sequencing Consortium, 2006; Wallberg et al., 2019). Indeed, to mitigate assembly problems related to repeated DNA and gene families, polymorphisms in the individual selected for sequencing must be as low as possible. In diploid species, this is addressed by selecting a highly inbred sample. Honey bees, however, have a haplodiploid sex-determination system, so the problem with intra-individual polymorphism is eliminated by sequencing a single haploid male (drone) (Figure 4(B)).

3.3.2. Annotating the reference genome

Much of the information used to annotate genes within reference genomes is derived from RNA-Seq gene expression studies, most of which are conducted using short-read sequencing. The annotation for new RefSeq genomes can be requested out at no cost for the user through the NCBI team, using the Eukaryotic annotation pipeline Gnomon (Thibaud-Nissen et al., 2013). However, short reads do not capture all the information, especially in large, complex genomes which may have highly repetitive sequences. Some genes may be falsely merged or split, exons can be missing, and overall, the information on alternative splicing is missing. Ideally, annotation-driven work should now be revised and performed with long reads such as Iso-seq (PacBio). Many tools computational tools exist for user to perform or curate their own annotation (Ejigu & Jung, 2020). New annotation pipelines user-friendly like BRAKER3 (Gabriel et al., 2024) are now using long read RNA-Seq and protein data to improve gene prediction.

3.3.3. High molecular weight DNA extraction

HAv3.1, being based on the DH4 honey bee strain, cannot represent the full diversity existing in all honey bee subspecies. Therefore, de novo assemblies for other representative individuals may need to be produced. Regardless of the DNA sequencing and mapping techniques used, a critical step is to ensure the genomic DNA extracted is of high molecular weight (HMW), typically 10 kb or higher. This is particularly important for long-read sequencing since

reads will be long only if the DNA molecular weight is sufficiently high. Achieving this requires careful sample handling and DNA extraction, which can be challenging at first. Most problems reside at the sampling stage and first stages of DNA extraction.

To prevent quality issues related to contaminants such as cuticle and optical pigments, samples should be collected from larvae or white-eyed pupal stage. Co-purification of pigments with HMW DNA can lead to errors in spectrophotometric measurements and interfere with downstream DNA binding, as observed in other organisms (Adema, 2021; Fauchery et al., 2023). To preserve tissue integrity, these should either be fresh or flash-frozen in liquid nitrogen. To preserve high molecular weight DNA, the frozen sample can either be pulverized and the powder used for standard DNA extraction. Alternatively, a freshly sampled or thawed tissue can be gently ground in DNA extraction buffer. While there are many options for mechanical disruption, using a pestle is ideal to avoid fragmenting the DNA. DNA extraction can then be performed using standard procedures, for instance, the QIAGEN Genomic-tips 100/G (Cat No/ID: 10243). The actual sequencing and genome assembly is a very specialized work, using ever-changing sequencing technology and bioinformatics pipelines (Childers et al., 2021; LaFlamme, 2021; Lawniczak et al., 2022; Rice & Green, 2019).

3.4. Small and large variant detection

The technology most often used for small and large variant detection (SNPs and indels, respectively) is Illumina, due to its high throughput, cost-effectiveness, and low error rate. When short reads are produced for each individual and aligned on the reference sequence, sequence differences between the reads and the reference can be observed, indicating the presence of variants. The bioinformatics analyses concerning SNP detection are described further in Section 4.3. Small variant detection can also be performed by pool sequencing, in which DNA from multiple samples are mixed prior to indexing and sequencing. Pool sequencing helps in reducing library preparation time and sequencing cost for estimating allele frequencies in populations.

To detect large variants, such as insertions or deletions of several hundred or thousands of base pairs (indels), either two de novo assembled genomes will have to be compared, as described further in Section 4.4 using the software LAST (Frith & Kawaguchi, 2015), or long reads produced by PacBio or Nanopore sequencing are aligned onto the reference for breakpoint detection. A common alternative to using LAST is minimap (Heng Li, 2018) or LASTZ (Harris, 2007).

3.4.1. RAD-seq

Restriction site-associated DNA sequencing (RAD-seq) is a technology by which a reduced representation of the genome is selected using restriction enzymes and PCR. Typically, 10% of the genome is selected for analyses to reduce the volume of sequencing reads required for SNP detection. However, given the small size of the honey bee genome, the number of reads obtained in an Illumina sequencing run is not the limiting factor. From our experience, using RAD-seq in honey bees now will increase the cost of library preparation and, by extension, the overall cost of sequencing while producing data for only a fraction of the genome. Moreover, bioinformatics analyses are slightly more complex than whole genome sequencing, and polymorphisms within the restriction sites used can cause allelic drop-out.

3.5. Sequencing museum specimens

Ancient and historic specimens offer an excellent and powerful opportunity to study macro- or micro-evolutionary changes (Short et al., 2018). Museum collections, in particular, are ideal, as they can provide a temporal series of honey bee samples from different areas of their natural range. Analyzing old specimens enables us to gain insight into past genetic diversity, selection, domestication, speciation, migration, and phylogenomics (Card et al., 2021; Raxworthy & Smith, 2021). In fact, to study these processes directly, comparing historic and contemporary allele frequencies is the most direct and powerful method, in contrast to model-based approaches.

The main caveat of old samples is the challenge in obtaining high-quality DNA for molecular analyses, but improvements in DNA extraction protocols and sequencing technology are overcoming these difficulties (Orlando et al., 2021). Until recently, most of the studies using museum collections were based on PCR-amplification of target genes or mitochondrial DNA. However, due to DNA degradation, fragments may be shorter than the target region and cannot be amplified. In contrast, high-throughput sequencing enables us to sequence even very short DNA fragments. In the field of human genetics, protocols for high-throughput sequencing applied to historic and ancient DNA from archeological sites are relatively well established, but fewer efforts have been made in the application to historical museum collections from animals (Card et al., 2021; Grewe et al., 2021; Mikheyev et al., 2015, 2017) and plants (Kistler et al., 2020). For honey bees, sequencing of historic collections has already been successfully performed to study signatures of selection, changes in ancestry composition, and genetic diversity pre- and post-parasite arrival (Cridland et al., 2018; Mikheyev

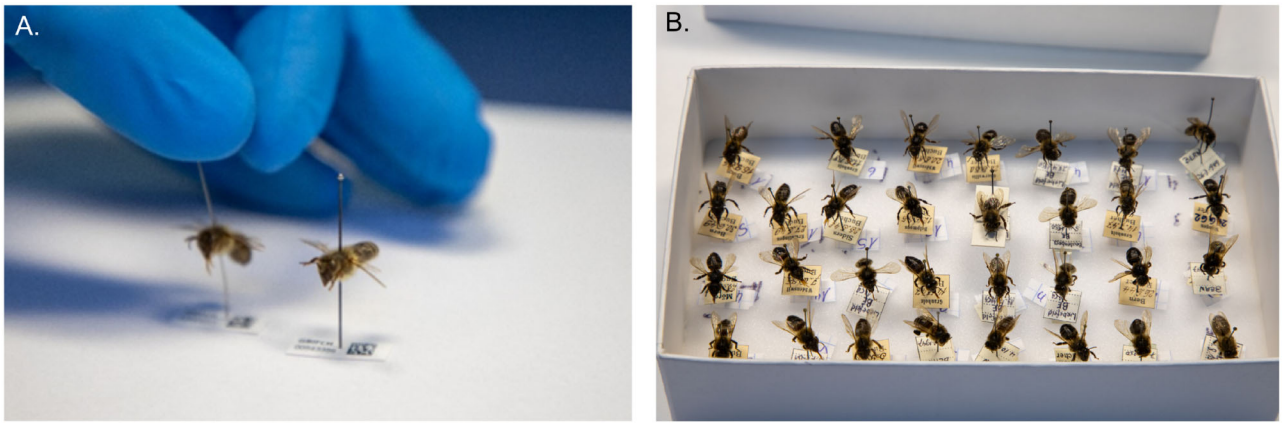

Figure 5. Examples of museum specimens. To avoid cross-contamination with other samples, especially with present-day specimens, dedicated rooms, special personal protective equipment, and a cleaning workflow must be adopted. (A) Single dried, pinned honey bee worker museum specimens. (B) Collection box with multiple museum specimens. (Credit: M. Parejo).

et al., 2015; Parejo et al., 2020) (Figure 5). The approaches employed are also applicable to other *Apis* species.

3.5.1. Considerations

The following recommendations are not limited to ancient or museum-grade specimens. We suggest adhering to similar guidelines for handling *Apis* samples targeted for genomic material extraction to reduce contamination.

- Reducing contamination: It is imperative to take the highest precautions and implement stringent measures to avoid any cross-contamination between samples, especially between contemporary and historical samples. We recommend strict separate handling of old and new samples, working under a hood or PCR workstation-type cleaned with UV light. The usage of filter tips for all steps should be seen as mandatory. Reliable DNA extraction from museum samples requires a clean and separated space where no previous bee material have been introduced. It is almost certain that such details will be demanded, and quality checked by reviewers of *Apis* museum genomics studies.
- Method choice: The DNA extraction protocol described below is based on phenol-chloroform separation, as it typically yields the highest amount of DNA. This is crucial because only limited tissues may be available, such as legs. Whenever enough tissue is available and samples are less degraded, commercial DNA extraction kits (e.g., DNeasy Blood and Tissue kit from QIAGEN on honey bees from 1910 to 1999) also produce adequate DNA quantity and quality (Cridland et al., 2018).
- Hazardous agent: Please note that phenol and chloroform are hazardous agents. It is of utmost importance that you work cleanly and safely. In addition to following basic laboratory safety precautions, perform all procedures in a chemical fume hood, wash hands thoroughly immediately after working with these chemicals, and use sealed safety tubes (e.g., Eppendorf Safe-Lock tube) when centrifuging.
- Material choice: The bulk tissue material used influences the success of DNA extraction and sequencing in chitinous organisms such as *Apis* honey bees. In some cases, there is no choice, as only non-destructive sampling is allowed by the curator for valuable museum specimens. While thorax sampling is typical for modern samples, thoraxes may contain a larger fraction of non-target DNA, such as contaminating fungi or bacteria (albeit far less than found in bee abdomens), and thus not represent the most suitable sample type. In contrast, hind legs have been shown to contain sufficient honey bee DNA with limited contamination (Parejo et al., 2020).
- Protocol optimization: We recommend conducting initial protocol testing on less valuable samples, if available, and changing or adapting the protocols accordingly. *Apis* honey bee species vary in body size while the anatomical structure remains similar. Thus, it is easy to understand the yield difference of ancient DNA obtained from a leg of the giant honey bee A. dorsata compared to a leg of the small A. florea honey bee. Not only species, but also the conditions of collection and preservation will affect the quantity and quality of the DNA fragments. Remember that a large proportion of your samples will not yield enough DNA of sufficient quality for subsequent sequencing.
- DNA degradation: It is not recommended to use a vortex to mix samples, as it can mechanically degrade DNA. Instead, a gentle and careful manual inversion of the tube several times to mix its content is advised.

3.5.2. Materials

1. Large equipment: thermoblock and cooling benchtop centrifuge
2. Dissection tools (cleaned with bleach and rinsed with ethanol)
3. Pipettes
4. Microtubes (1.5 and 2 ml, Eppendorf)
5. Saline solution for initial cleaning
   - Ringer solution (0.125M sodium chloride, 1.5mM calcium chloride dihydrate, 5mM potassium chloride, 0.8mM sodium phosphate dibasic, pH 7.4)
   - Alternatively, PBS (137 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4, and 1.8 mM KH2PO4, pH 7.4) or Tris solution (150 mM NaCl, 50 mM Tris-HCl, pH 7.6) can be used
     - Lysis buffer: ATL (QIAGEN or alternatively other commercial lysis buffers can be used)
     - Proteinase K (20 mg/ml)
     - Phenol-chloroform-isoamyl alcohol (25:24:1)
     - Chloroform
     - Molecular grade glycogen (10 mg/ml)
     - Ethanol (72 and 100% solutions)
     - Sodium acetate (0.3 M final concentration, pH 5.2)
     - Ultra pure water (Milli-Q)

3.5.3. Procedure

3.5.3.1. Preparation and lysis

1. Clean 3–10% bench surface with bleach and rinse with 72% ethanol.
2. Take the museum specimen, carefully cut the hind leg(s) and place them into a 1.5–2 ml Eppendorf tube filled with 1,000 µl Ringer solution.
3. Gently invert for 20 min to clean and rehydrate.
4. Take out the leg(s) and place on a new paper tissue to dry.
5. Add 180 µl lysis buffer to a new 1.5 ml Eppendorf tube.
6. Add leg(s).
7. Add 20 µl of proteinase K, invert and quick spin.
8. Incubate overnight at 56 °C.

3.5.3.2. DNA extraction

9. After incubation, centrifuge and quick spin.
10. Transfer supernatant (~190 µl) to a new tube (leaving the solid part of the tissue at the bottom).
11. Add 190 µl (equal amount) of phenol:chloroform:isoamyl alcohol to the supernatant.
12. Invert carefully for 5 min.
13. Centrifuge at 12,000 g for 5 min.

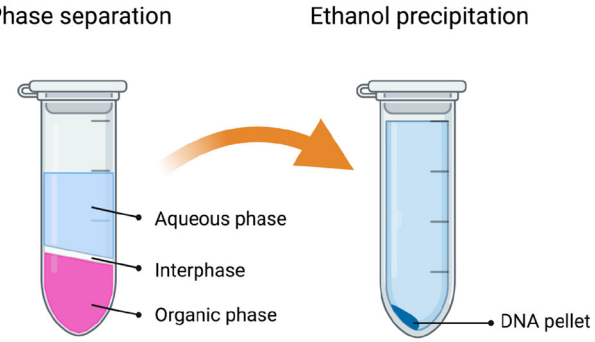

Figure 6. Sample appearance during phase separation and precipitation steps during the extraction. Created with Biorender.com.

14. Pipette upper aqueous phase (~100 µl) into a new tube, making sure not to touch the interphase (Figure 6).
15. Add 190 µl H_2O (Milli-Q) to the remaining interphase and organic phase (back extraction).
16. Invert carefully for 5 min.
17. Centrifuge at 12,000 g for 5 min.
18. Pipette upper aqueous phase (~200 µl) into the same tube containing previously pipetted aqueous phase, making sure not to touch the interphase.
19. Add 300 µl of chloroform to the tube with aqueous phase (equal volume) to remove remaining phenol.
20. Invert carefully for 5 min.
21. Centrifuge at 12,000 g for 5 min.
22. Pipette supernatant into a new tube, making sure not to touch the interphase.

3.5.3.3. Precipitation

23. Add 1 µl of 10 mg/ml glycogen solution to help make the pellet visible upon precipitation.
24. Invert for 2 min.
25. Add 20 µl sodium acetate (10% of the volume).
26. Add 2.5–3 volumes of ice-cold 100% ethanol (~600 µl).
27. Place the tube at −80 °C for 1.5–2 h.

3.5.3.4. Washing

28. Centrifuge at 13,000 g at 4 °C for 30 min.
29. Identify the pellet, and pipette away liquid.
30. Add 850 µl 72% ice-cold ethanol.
31. Invert carefully 1–3 times making sure not to detach the pellet.
32. Centrifuge at 13,000 g at 4 °C for 10 min.
33. Identify the pellet, and pipette away liquid.
34. Repeat the wash steps 30–33.
35. Quick spin.
36. Pipette away the last drops of ethanol and let it dry for 30 min at 37 °C to let it dry.

3.5.3.5. Final solubilization

37. Add 40 µl of H_2O (Milli-Q).
38. Carefully mix by pipetting up and down several times (do not vortex).
39. Incubate at 4 °C overnight for full solubilization.

TIP: A final, optional DNA purification step can be performed using magnetic beads (e.g., QIAGEN's EZ1® DNA Tissue Kit). This step will remove co-extracted small molecules that could act as inhibitors in downstream enzymatic reactions. However, it will also yield less total extracted DNA.

3.5.4. Sequencing of museum and ancient genomes

Depending on the age and storage conditions of your samples, the DNA will be more or less degraded and different subsequent sequencing strategies might be used. If sufficiently high-quality DNA is extracted, standard whole-genome sequencing libraries can be prepared (Section 4). In fact, most of the historic and ancient DNA sequence data have been produced using Illumina technology due to the high data output, cost-effectiveness, and relatively low error rates. It works well on short sequences and is thus particularly well suited to sequence DNA in the range of 50–150 bp, which characterizes most degraded old DNA. Specific protocols for library preparation of ultra-low quantity DNA exist from a variety of companies (e.g., Nextera DNA Library Preparation Kit (Illumina); NEBNext Ultra II kit (New England Biolabs, Inc). For degraded DNA it is not recommended to perform any size selection during the library preparation. We recommend to check with your local sequencing facility for recommendations. Before sequencing, we suggest to evaluate degradation levels and identify possible pollutants, such as formaldehyde, in the prepared libraries with a fragment analyzer (e.g., Bioanalyzer or 4200 TapeStation System (Agilent Technologies Inc.)).

As an alternative to whole-genome shot-gun sequencing, it is also possible to utilize hybridization capture methods. In particular, for samples with low endogenous DNA content, shot-gun sequencing is inefficient and expensive. For such samples, hybridization enrichment provides an approach to enrich a DNA sample for specific (and larger) genomic regions (for example, targeted genes, the exome or mitogenome). These approaches not only reduce analytical costs but also maximize the chances of identifying DNA present even in limited abundance. A typical limitation of hybridization-based genome reduction techniques is that custom probe design and synthesis are rather laborious and costly. Recently, a new and cost-effective method, hybridization restriction-associated-DNA (HyRAD) has been developed that uses double enzymatic restriction of fresh DNA extracts to build RNA probes that cover only a fraction of the genome and can serve as baits for capturing homologous fragments from old DNA libraries (Suchan et al., 2021). This can be a suitable, cost-effective approach for non-model organisms, such as *Apis* species.

For *A. mellifera*, which has well-established genomic resources, an alternative approach is to use SNP panels. These panels target specific SNPs that can provide information on population structure, diversity and ancestry, or can be used to investigate functional genes or alleles that may be associated with particular phenotypes. Custom genotyping panels can be created, but a number of pre-designed SNP arrays are available as ready-to-use assays for the western honey bee (Chapman et al., 2017, 2015; Henriques et al., 2021; Momeni et al., 2021; Muñoz et al., 2015). A caveat to keep in mind for both targeted enrichment and SNP panel approaches are the ascertainment bias that can be introduced by the selection of particular probe sets, which can affect downstream estimates of population genetic statistics.

Finally, novel sequencing techniques are constantly being developed. We recommend following the latest research of human ancient genomics for an overview on the latest developments (Orlando et al., 2021).

3.5.5. Guidance on the data analysis methods

Sequence data clean-up and strict filtering is crucial for reliable variant calling and subsequent analyses and interpretation of the data. Ancient DNA has "mistakes" at the ends caused by deamination of cytosine, which are converted into uracil and thereafter sequenced as thymine analogues. This process is responsible for the sequencing artifacts observed as misincorporations of G→A at the 5′ and C→T at the 3′ ends. The frequency of this DNA damage increases with the age of a sample and can be statistically evaluated and taken into account during bioinformatic analysis. To evaluate DNA damage, it is recommended to perform a quality check of raw sequence reads using (e.g., DamageProfiler or mapDamage2. The latter program also enables users to rescale quality scores of likely damaged positions in the reads. In this way, "damaged" positions will have less weight for mapping and subsequent variant calling. A critical step is to use strict read trimming to remove adapters and low-quality terminal sequences (as identified for instance by DamageProfiler). Also overlapping read pairs, resulting from shorter DNA fragments than sequencing read length, may be collapsed into consensus sequences (e.g., by using BBmerge).

Once the raw sequence data has been sufficiently cleaned, standard variant calling pipelines and analyses can be followed (see Sections 3.4 and 4.3). Some tips include adjusting parameters for mapping specificity and sensitivity. Moreover, rather than actual variant calling, it is possible to use genotype likelihoods instead, and in this way account for the uncertainties and lower base qualities of ancient DNA bases. This can be done within the framework of The Genome Analysis Toolkit (GATK) or Analysis of Next Generation Sequencing Data (ANGSD) that covers a wide range of analytical population genetics statistics.

3.5.6. Applications and limitations

Any ancient historic DNA study is only as good as the sampling and careful interpretation of results. For instance, low genetic diversity in historic samples could result from the fact that all honey bees were from the same colony. Often in ancient and museum specimens, limited knowledge of sampling conditions is available, and it may not be known whether the samples were collected from managed hives or in the wild. Cautious interpretation is thus crucial. Nevertheless, ancient honey bee genomics of museum samples has a great potential to illuminate the evolutionary history of species and different subspecies, to identify the timing of population splits, and investigate signatures of natural selection (Cridland et al., 2018; Mikheyev et al., 2015; Parejo et al., 2020). Moreover, the approaches can not only be applied to study the past diversity of bees but also their historically associated pathogens and microbiomes. Thus, new lines of investigation could be the study of the evolution and spread of pathogens and the genetic responses of hosts, as well as past bee health related to its microbiome.

4. Whole-genome population and association studies

4.1. Introduction

When considering whole-genome analyses, one major advantage of working with *A. mellifera*, *A. cerana* or even *A. dorsata* is that their genomes are very compact, only ~223–228 Mb long (Oppenheim et al., 2020; Wallberg et al., 2019; Wang et al., 2020). In comparison, the honey bee parasite Varroa harvest a larger genome size of 368 Mb (Cornman et al., 2010; Techer et al., 2019) while the human genome, for instance, is even larger with 3055 Mb (Nurk et al., 2022). As a result, it quickly became evident that the cost-to-benefit ratio of using whole genome sequencing for honey bee applications was excellent, as compared to other approaches. For instance, SNP chips only allow the study of between 10,000 and 100,000 markers for a price that will not be much lower than whole genome sequencing (between one third and half the price at the time of writing), whereas the latter approach enables the detection of several million markers. Other disadvantages of using SNP chips are that the choice of markers is biased, and that the high density of SNPs and indels in the *A. mellifera* genome complicates the chip design (allele drop-out can often happen due to neighboring polymorphism).

Nevertheless, the whole-genome sequencing approach has its drawbacks, with complex and intricate analysis pipelines and longer computational times. While bioinformatics is included in most molecular biology and ecology related curriculum, the steep learning curve can be discouraging for new users without proper guidance (Carvalho & Rustici, 2013). Additionally, whole genome computations should really be carried on remote access high performance clusters (HPC) or local workstations. It may thus not be ideal if rapid answers to questions such as parentage testing or subspecies assignment are needed. Moreover, data storage can be a critical issue, with an average of 2–3 Gb needed per sample for the raw sequencing files (FASTQ: text-based file that contains the nucleotide sequence and associated quality score), 3–4 Gb for the corresponding alignment files (BAM: binary alignment map files) and files up to one terabyte for reporting genomic sequence variation (VCF: variant calling files) depending on the dataset size.

Until now, most sequencing results were obtained with paired-end Illumina sequencing. With the recent progress made in long-read sequencing, having reference-quality assemblies for several individuals, ideally from different subspecies within a targeted *Apis* species can become a future standard. This will allow the construction of pan-genome graphs, taking inter-individual variation into account at the reference genome level. Long-read sequencing of individual samples will also allow the detection of large structural rearrangements and repeat landscapes that cannot be analysed accurately with short reads.

In this chapter, we provide recommendations for multiple types of software to analyze genome-wide sequencing data. While we propose a bioinformatics workflow, it is important to note that this field is rapidly evolving, and new and improved tools may emerge over time. Before using one software, users should verify its maintenance status (e.g., version, last update, accessibility to a forum, or troubleshooting documentation) to avoid using obsolete or buggy versions (Section 11). To carry out most bioinformatics tasks, a basic understanding of UNIX shell language (bash) is necessary on terminal and non-graphical friendly interfaces (cheat sheet

recommended for new users). Languages such as R, Python, awk, and Perl are worth mastering for bioinformatics-oriented researchers (although Python is now generally preferred over Perl, current bioinformatic pipelines still utilize scripts in both languages).

4.2. Ploidy and sampling considerations

Since male bees are haploid and female bees are diploid, the genetic information derived from these samples will be different, and the best sample type will vary depending on the study's goals. Additionally, sampling individuals versus groups of individuals will change the types of questions that can be answered with the resulting genomic data. Below, we outline some information to help determine what type of sample to work with.

4.2.1. Individual sampling

Genotypic information for workers and queens match standard polymorphism data showing three genotypes: homozygous for the reference allele, heterozygous, and homozygous for the alternative allele. Drones, being haploid, only two genotypes: the reference or the alternative allele. Workers provide twice as much data as drones due to their diploid nature, offering better insights into population structure and admixture. However, drone analysis requires less sequencing coverage and is more cost-effective.

4.2.2. Pooled sampling

In addition to individual samples, depending on the goal of the experiment, pooled sampling may be necessary. Since workers carry both queen's and inseminating males' genetics, a pooled worker sample may capture the entire allelic diversity within a colony at a more affordable cost than individual analysis. In this case, 50% of the pooled workers' genetics are derived from the queen, whereas the remaining 50% is representative of the patrilines (the cohort of drones that inseminated the queen) of the colony (see Section 4.3.7 for a brief discussion on handling pooled data). Pools can be made from honey bee tissue (e.g., thoraces, flight muscles, heads with compound eyes removed, legs, antennae) or from multiple specimen DNA.

4.2.2.1. Groups of workers. To represent the complete genetic makeup of the meta organism that is a honey bee colony, a large number of individuals should be sampled. This requirement is a consequence of the polyandrous mating system in *Apis* bees with variable degrees across species (Kraus et al., 2005; Palmer & Oldroyd, 2000). While extreme polyandry levels were reported in the giant honey bee *A. dorsata* (up to 102 detected patrilines (Wattanachaiyingcharoen et al., 2003)) the following advices relates *A. mellifera* mating frequency (averaging around 13.8 ± 2.5 males (Palmer & Oldroyd, 2000)). The number of pooled individuals has ranged from $n = 12$ to 500 (Guichard et al., 2021; Momeni et al., 2021; Rizwan et al., 2020; Saelao et al., 2020), and we recommend sampling and sequencing in the order of hundreds of individuals. For example, if 15 males inseminated the queen and a targeted sequencing depth is 30-fold per patriline is desired, pooling approximately 500 related workers would be necessary. Each worker in the pool would then represent 1/15th of the male genetic contribution of the colony, assuming equal representation of all patrilines.

4.2.2.2. Groups of drones. Alternatively, a group of drones can also be sampled. Pooling drones into the same genotyping experiment will allow indirect inference of the queen's genetic information, as drones carry the queen's genetics exclusively. However, to obtain a complete representation of the queen's genetics, it is still necessary to pool a significant number of drone offspring together (Jones, Du, et al., 2020).

4.3. SNP and indel detection

Here, we provide a primer on the bioinformatics involved in detecting sequence variants from Illumina data, as it is often very difficult to know how to start. The software tools chosen here result from the authors' positive experience in their application of *A. mellifera* genomics workflow but do not reject the validity and rapidity of alternative algorithms and pipelines. We recommend the review by Bourgeois and Warren (2021) listing the most common software and methods applicable to comparative and population genomics in eukaryotes. For further details on each parameter, future users will have to go to the dedicated websites for BWA-MEM, SAMtools, GATK, Picard, VCFtools, and BCFtools (Danecek et al., 2011, 2021; McKenna et al., 2010).

4.3.1. Mapping reads with BWA-MEM: from FASTQ files to BAM files

The first step is to map the short reads to the reference genome to produce a BAM file. Typically, for SNP detection, sequencing is performed paired-end with an Illumina instrument, giving millions of ~150 bp reads in which 300–500 bp will separate the beginning of Read 1 (first mate or forward read) and the end of Read 2 (second mate or reverse read), according to the size specification of the sequencing library. For each sample, the sequencing instrument

will produce two files in compressed FASTQ format containing the reads. These will have names in the form "AOC5_TAGCTT_L002_R1.fastq.gz" and "AOC5_TAGCTT_L002_R2.fastq.gz," which includes the sample name (AOC5), the sample identification in the sample sheet (TAGCTT), the lane number on the sequencing machine (L002) and the Read (R1 or R2). One file will contain all Read 1 sequences and another will contain all Read 2 sequences.

In honey bee genomics, there is currently no consensus on a single best mapping software. However, based on our literature survey, we found that **BWA-MEM** (Burrows–Wheeler Aligner), **Bowtie2** and Stampy have been used for *A. mellifera* and *A. cerana* short DNA read mapping. However, we noted that **BWA-MEM** was the most commonly used, and it was also reported as one of the top performing short reads aligners (Musich et al., 2021). Therefore, we have chosen BWA-MEM as our current standard software for mapping honey bee genomes.

Step 1. Mapping the reads to the reference genome is done here with **BWA-MEM** (Li, 2013) and parsing with **SAMtools** using the following commands:

```
bwa mem -M GCF_003254395.2_Amel_HAv3.1_genomic.fna \
AOC5_TAGCTT_L002_R1.fastq.gz \
AOC5_TAGCTT_L002_R2.fastq.gz | \
samtools view -bh -o AOC5_TAGCTT_L002_aligned.bam -
```

The `-M` command marks shorter split reads as secondary. As **BWA-MEM** will output a very large file in SAM format, its output is piped (|) directly into the `-samtools view -bh` command, that will compress directly in the compact BAM format (`-b`), while including the header (`-h`).

The full description of the SAM format can be found online, and is updated on a regular basis. Briefly, a SAM/BAM file will contain a header with diverse information, such as if the file was sorted, the name of the reference sequence used, and the programs that were used to process the data (alignment and other subsequent analyses).

Special considerations: If a sample was sequenced in two separate runs, lanes or from two different libraries, the best practice is to require the mapping to be done separately and the BAM files produced to be subsequently merged. This is important for backwards traceability and for downstream analyses, as reads from different runs may have specific biases affecting base or mapping quality estimations. In this scenario, sample "BER15" was sequenced in two separate lanes of an Illumina instrument. Alignment was done separately, giving two BAM files, to be merged with `samtools merge`.

Step 2. Merge BAM files with mapped reads using SAMtools:

```
samtools merge -o BER15_ATCACG_merged_aligned.bam \
BER15_ATCACG_L002_aligned.bam \
BER15_ATCACG_L005_aligned.bam
```

This information is passed on to the BAM file in the mandatory ID tag of the read group (@RG) lines in the header, and each mate-pair read in the BAM file will be assigned to one or the other ID tag:

```
samtools view -H BER15_ATCACG_merged_aligned.bam | grep ^@RG
@RG ID:BER15_ATCACG_L002
@RG ID:BER15_ATCACG_L005
```

Step 3. Sort the BAM file using the following commands with Picard:

```
java -jar picard.jar SortSam \
      I = AOC5_TAGCTT_L002_aligned.bam \
      O = AOC5_TAGCTT_L002_sorted.bam \
      SORT_ORDER = coordinate
```

4.3.2. Marking duplicate reads with Picard

During sample preparation and sequencing, DNA amplification steps will be performed, in which case a single DNA fragment can produce duplicate reads. It is important to tag such reads in the BAM file, so that they will be ignored when calling variants with GATK `-HaplotypeCaller`.

Step 4. Tag the duplicate reads using Picard option `-MarkDuplicates`:

```
java -jar picard.jar MarkDuplicates \
      I = AOC5_TAGCTT_L002_sorted.bam \
      O = AOC5_TAGCTT_L002_dedup.bam \
      M = marked_dup_metrics.txt
```

Step 5. The reads can then be sorted and indexed with **SAMtools:**

```
samtools sort -T AOC5_TAGCTT_L002_dedup.bam \
-o AOC5_TAGCTT_L002_sort.bam
samtools index AOC5_TAGCTT_L002_sort.bam
```

4.3.3. Base quality score recalibration (BQSR) with GATK

Sequence reads in the FASTQ format have a quality score associated to each base called by the sequencing machine, which are based on the manufacturer's algorithms. These express the confidence of the base calling and will greatly influence the algorithms used for variant calling as well as deciding between sequencing errors and real biological differences.

BQSR models non-random technical errors in the data using a machine learning approach and adjusts the scores accordingly. A set of known variants in the VCF file format is provided to mask bases at sites of expected variation. Mismatches outside these sites are counted as errors.

Step 6. Use the following GATK commands to generate the recalibration table:

```
gatk BaseRecalibrator \
   -I AOC5_TAGCTT_L002_sort.bam \
   -R GCF_003254395.2_Amel_Hav3.1_genomic.fna \
   --known-sites sites_of_variation.vcf \
   --known-sites another/optional/setOfSitesToMask.vcf \
   -O recal_data.table
```

Step 7. Recalibrate base qualities:

```
gatk ApplyBQSR \
   -R GCF_003254395.2_Amel_Hav3.1_genomic.fna \
   -I AOC5_TAGCTT_L002_sort.bam \
   --bqsr-recal-file recal_data.table \
   -O AOC5_TAGCTT_L002_BQSR.bam
```

4.3.4. Calling variants with GATK

Step 8. Variants can be called for each sample individually using the following GATK commands:

```
gatk HaplotypeCaller \
   -R GCF_003254395.2_Amel_Hav3.1_genomic.fna \
   -I AOC5_TAGCTT_L002_BQSR.bam \
   -O AOC5.g.vcf.gz
   -ERC GVCF \
   -ploidy 2
```

Special considerations: The option "`-ploidy`" is set to 2 (diploid) in the example, which fits for a honey bee worker or queen sample. The option "`-ploidy 1`" can be used for haploid drones, but in our experience, this leads to false positive SNP detection. A better option is to analyze haploid samples using the diploid model and to filter out SNPs with heterozygote calls in the drones after the genotyping step.

4.3.5. Combining all samples and genotypes with GATK

At this stage, all individual genotype files will be combined into a single file. If more than hundred samples are processed, this may require a large amount of memory, which can be specified (`--java-options "-Xmx10g"`).

Step 9. Combine genotype files:

```
gatk --java-options "-Xmx10g" CombineGVCFs \
   -R GCF_003254395.2_Amel_Hav3.1_genomic.fna \
   --variant AOC1.g.vcf.gz \
   --variant AOC2.g.vcf.gz \
   --variant AOC3.g.vcf.gz \
   --variant AOC4.g.vcf.gz \
   --variant AOC5.g.vcf.gz \
   -o all_samples.g.vcf.gz
```

Step 10. Once combined in a large file, joint genotyping can be performed:

```
gatk --java-options "-Xmx10g" GenotypeGVCFs \
   -R GCF_003254395.2_Amel_Hav3.1_genomic.fna \
   -V all_samples.g.vcf.gz \
   --use-new-qual-calculator \
   -O all_samples_genotyped.vcf.gz
```

4.3.6. Filtering variants with GATK: technical filters

Step 11. The following commands will retain only site variants coded as SNPs:

```
gatk SelectVariants \
   -R GCF_003254395.2_Amel_Hav3.1_genomic.fna \
   -V all_samples_genotyped.vcf.gz \
   --select-type-to-include SNP \
   -O all_samples_genotyped_SNPs.vcf.gz
```

Step 12. Variants can then be filtered based on quality scores. First, mark the SNPs in the VCF file that do not pass quality score thresholds:

```
gatk VariantFiltration \
   -R GCF_003254395.2_Amel_Hav3.1_genomic.fna \
   --filter-expression « QD < 2.0 || FS > 60.0 || MQ < 40.0 »
   --filterName "quality_filters"
   -V all_samples_genotyped_SNPs.vcf.gz \
   -O all_samples_genotyped_SNPs_filter_tagged.vcf.gz
```

Here, we adopted the recommended and generic hard-filtering parameters of all SNP with a variant confidence (QD) below 2, a strand bias (FS) over 60 and a mapping quality (MQ) below 40. However, we recommend readers to always plot these statistics distribution using R and ggplot2 and verify that the thresholds are adapted to their dataset.

Step 13. Then, the SNPs that passed the filters can be extracted:

```
gatk SelectVariants \
   -R GCF_003254395.2_Amel_Hav3.1_genomic.fna \
   --exclude-filtered
   -V all_samples_genotyped_SNPs_filter_tagged.vcf.gz
   -O all_samples_genotyped_SNPs_filter_passed.vcf.gz
```

Special considerations: If variants were detected in haploid drones using the diploid model, markers having heterozygote calls can now be removed.

4.3.7. Filtering variants with VCFtools: data quality

To ensure a high-quality dataset, SNPs should be filtered using not only quality scores, but also, for instance, the mean depth and missing data. Here, the software VCFtools (Danecek et al., 2011) is used to include biallelic SNPs (`--max-alleles`) with at least 5 read depth (`--min-meanDP`). Further information about SNP filtering, such as the correlation between markers or linkage disequilibrium (LD), can be found in Section 4.7.

```
vcftools \
--gzvcf all_samples_genotyped_SNPs_filter_passed.vcf.gz \
--max-alleles 2 \
--min-meanDP 5 \
--out all_samples_filter_2.vcf.gz
```

4.3.8. Genotype phasing

Phasing refers to the process of statistical estimation of haplotypes from genotype data. While drone

sequencing generates directly phased data, worker or queen sequencing requires an extra step in the analysis for phasing the diploid data. There are several freely available programs for this purpose, such as **Beagle 5.4** (Browning et al., 2021), **fastPHASE** (Scheet & Stephens, 2006), **IMPUTE2** (Howie et al., 2009), and **SHAPEIT4** (Delaneau et al., 2019). **SHAPEIT4** is highly accurate, simple, and computationally fast in large datasets (De Marino et al., 2022). The program can be implemented using the following command line:

```
shapeit4.2 - -input all_samples_filter_2.vcf.gz \
- -map genomic_map.gmap.gz \
- -output phased.vcf.gz
```

4.3.9. SNP annotation with SnpEff

SnpEff is a toolbox to annotate the variants and to calculate their effects (Cingolani et al., 2012). **SnpEff** can be implemented on the Galaxy analysis platform or by using the following command line:

```
java -Xmx8g -jar snpEff.jar Apis_mellifera all_samples_filter_
2.vcf.gz > all_samples_filter_2_annotated.vcf.gz
```

4.3.10. SNP analysis by sequencing pooled samples

Pooling offers an alternative for screening populations or subspecies using genome-wide markers and high-throughput technologies. Group-level data obtained from genotyping experiments will deviate from individual-level data in the sense that frequencies of the different alleles observed in the pool, rather than genotype polymorphisms, will be obtained (see Section 4.2.2). Allele frequencies can be used directly or transformed into mean genotype using **PLINK** (see Section 4.5.4.3), or dedicated genotype reconstruction methods. Such methods are currently available to use such allele frequencies, from drone offspring or from a worker pool (Eynard et al., 2022) to obtain queen genotypes. The procedure described here allows the estimation of allele frequencies in pooled samples by using the **PoPoolation** software (Kofler et al., 2011).

4.3.10.1. From FASTQ files to BAM files. Mapping reads and duplicate reads detection are performed as described above (Section 4.3.1 and Section 4.3.2).

4.3.10.2. SNP selection and pileup files. Each line in a PILEUP file describes a single position in the genome. It contains information on the number of reads "piled up" at the position, together with the base called for each read, if the reads mapped to the forward or the reverse strand and quality scores for the base called for each read. PILEUP files are very large and the detection of SNPs in pooled sequencing is not very effective, so it is best to work on a list of high-quality SNPs that were detected in a previous experiment.

```
samtools mpileup -I -l snp.list \
      -f reference_genome.fa \
      -C 50 -q 20 -Q 20 sample.bam \
      -o sample.pileup
```

The option `-I` is for skipping indels. The list `snp.list` is in the form of chromosome position (tab separated). This will generate one PILEUP file per sample. A list of BAM files can be provided (option `-l`), to make a mpileup file containing multiple samples.

```
samtools mpileup -I -l snp.list \
      -f reference_genome.fa \
      -C 50 -q 20 -Q 20 -b bams.list \
      -o multiple_samples.mpileup
```

4.3.10.3. Counting reads per allele with PoPoolation. PoPoolation (Kofler et al., 2011) takes PILEUP files as entries and generates read counts for each possible nucleotide base (A, C, G, or T) at each SNP position reported in the PILEUP file.

```
java -ea -Xmx10g -jar mpileup2sync.jar \
                 - -input sample.pileup \
                 - -output sample.sync \
                 - -fastq-type sanger
```

The `-sample.sync` files present the results in the form of one line per SNP position, with a count of reads for each of A:T:C:G:N:del. The **PoPoolation** suite provides some perl scripts to calculate allele frequency differences or F_{ST} (fixation index) values between pairs of pooled DNA sequences.

4.4. Comparing whole genomes

Comparing different versions of whole-genome assemblies can help assess their quality and identify major structural rearrangements occurring between individuals of the same species, from different subspecies, or different species. The rearrangements thus detected will be large deletions, insertions, inversions, or translocations. There are many software options for pairwise alignments, including **LAST** (Frith & Kawaguchi, 2015), **MUMmer** (Marçais et al., 2018) or, **minimap** (Heng Li, 2018), among others. These tools can also be used to align long-reads from *PacBio SMRT* (single-molecule real-time) sequencing technology or *Oxford Nanopore Technologies* (ONT) on reference genomes, thus enabling rearrangements to be detected directly after the assembly process. The case of simultaneous and multiple alignments is more complex and will not be described here (see currently developed methods

(Armstrong et al., 2020; Kille et al., 2022) and large eukaryote dataset applications (Feng et al., 2020; Zoonomia Consortium, 2020)).

4.4.1. Conducting a pairwise genome comparison with LAST

The example of pairwise whole-genome comparison given here, using LAST, describes the main steps with default values for the options proposed being used as much as possible. Detailed information on the numerous options proposed by the software that can affect sensitivity, speed and other aspects are described in the software documentation. In the following example, we will compare the AMelMel genome assembly of the black bee *A. mellifera* mellifera (query sequence), to the reference genome HAv3.1 (target sequence).

Step 1. Both the target (HAV3_1.fa) and query (AMelMel_1.fa) sequences must be in FASTA format, in the form:

```
cat AMelMel_1.fa
>Chromosome1
ACTACAGGATATCCATAGACAT …
>Chromosome2
GTCAGGATAGACAGGTAGACAT …
```

Use basic Unix commands such as `cat` or `less` to print the content of your `target.fa` and `query.fa`.

Step 2. Prepare index files for the target sequence with the command lastdb:

```
lastdb -uNEAR HAV3_1Db HAV3_1.fa
```

Note: The option `-uNEAR` specifies a seeding scheme that is good for finding strong similarities. It is used here instead of the default YASS seeding scheme, as we are comparing two subspecies and therefore very similar sequences.

Step 3. Define an optimal scoring matrix. This is optional, as **LAST** can work with a default matrix or with one of the matrices provided. The command `last-train` will find suitable substitution and gap scores for aligning the two sequences provided by using an iterative procedure:

```
last-train HAV3_1Db AMelMel_1.fa > score_matrix.mat
```

Step 4. Align the query to the target:

```
lastal -p score_matrix.mat HAV3_1Db AMelMel_1.fa > HAv_Amel.maf
```

Note: The option `-p score_matrix.mat` specifies the score matrix prepared with `last-train`. Any built-in score matrix proposed by LAST can be used.

Step 5. Plot the original alignment results. We will only plot the alignment of chromosomes 7 (NC_037644.1 in the HAv3.1 genome and CM010325.2 in the AMelMel genome) (Figure 7(A)). The next steps will aim at finding an unique best alignment.

```
last-dotplot -1 NC_037644.1 -2 CM010325.2 HAV3_1_AMelMel1_1.maf
plot_chr7.png
```

Step 6. Finding unique best hits for the query. At this point, any sequence segment in the query can have several alignments in the target. The command `last-split` will read the candidate alignments of the query sequences, and looks for a unique best alignment for each part of each query.

```
last-split HAV3_1_AMelMel1_1.maf > HAV3_1_AMelMel1_1_split1.maf
```

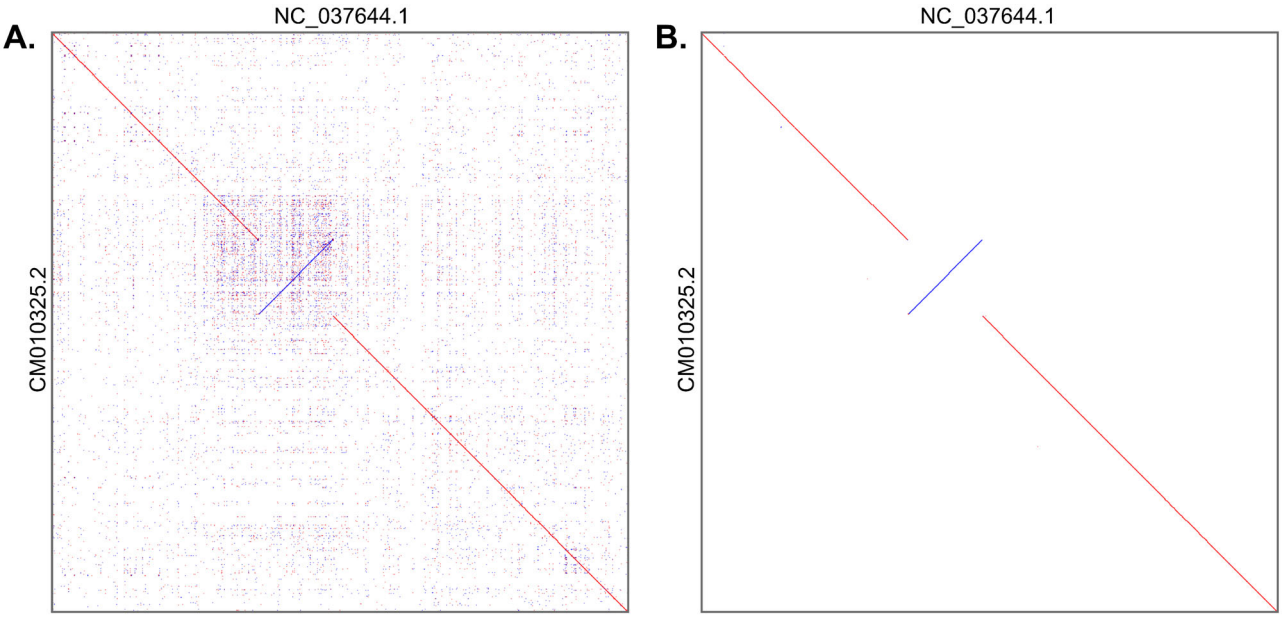

Figure 7. Example plots. (A) One major inversion is detected between the two assemblies. The many dots off the diagonal are due to sequences that are repeated and having therefore several reciprocal hits. (B) Each region of one genome is now aligned to at most one region of the other.

Step 7. Finding unique best hits for the target. As the sequence segments in the target HAv3.1 genome can also have several alignments in the query, query and target are inverted in the file, and `last-split` is run again:

```
maf-swap HAV3_1_AMelMel1_1_split1.maf | last-split > HAV3_1_
AMelMel1_1_split2.maf
```

Step 8. Plot the polished alignment results (Figure 7(B)).

```
last-dotplot -2 CM010325.2 -1 NC_037644.1 HAV3_1_AMelMel1_1_
split2.maf plot_chr7.png
```

4.5. Genome-wide association studies

Genome-wide association studies (GWAS) are conducted to detect genetic markers that contribute to phenotypic variation between individuals by analyzing phenotypic and genomic information in an unified statistical model (Figure 8). This type of analysis is commonly conducted in livestock species and plants, and an extensive literature and collection of tools, software, and methodologies are available for this purpose (for a review in animal genetics, see Hayes and Goddard (2010)). In this section, we will discuss GWAS in the context of honey bee studies, considering the unique attributes of this species.

4.5.1. Considerations for phenotypic data

Many different phenotypes or observations are compatible with GWAS. Examples for *Apis* honey bees include, but are not limited to:

- Traits of agronomic interest, such as honey production, gentleness, or low propensity to swarm
- Traits linked to phenotypic plasticity and social behavior, such as precocious forager age
- Traits linked to morphology, such as pigmentation or wing venation
- Traits linked to survival, such as resistance to parasites (Varroa spp., Tropilaelaps spp., small hive beetle), to viruses, or to specific environmental conditions
- Traits linked to the colony or individual health, such as viral or bacterial load
- Traits focusing on genome structure, such as variable recombination rate

Ideally, these characteristics should be measured in a standard way across all the "individuals" or "units" (e.g., colony, caste, or single bee specimen). Measuring traits at the superorganism unit (colony) or at individual levels can be more or less relevant depending on the trait of interest. Traits can be binary, as in a case/control association study (e.g., survival, presence or absence of a virus, performance or not of behavior, etc.), categorical (e.g., gentleness scored in classes), or continuous (e.g., honey production measured in kilograms). If focusing on a continuous phenotype, it is necessary to analyze individuals representing the whole expected spectrum to distinguish markers impacting the phenotype.

4.5.2. Considerations for sample selection

GWAS can be performed on individual samples or composite samples of individuals from the same colony (see Section 4.2 for details on ploidy considerations). Individual samples can represent different colonies or, when more than one individual per colony is sampled, individuals within a colony. These samples may be workers, queens, or drones. For any sample type, an appropriate sample size should be estimated to ensure the analysis has sufficient power to detect associations.

4.5.2.1. Power analysis. Prior to performing a case/control association study, it is useful to estimate the power needed to detect significant markers given a specific experimental design. Such power of detection is linked to the number of individuals in the case versus control group, the number of markers tested and the difference in allele frequency between the two groups. The R package pwr is strictly dedicated for power analysis functions (Champely et al., 2017, 2018).

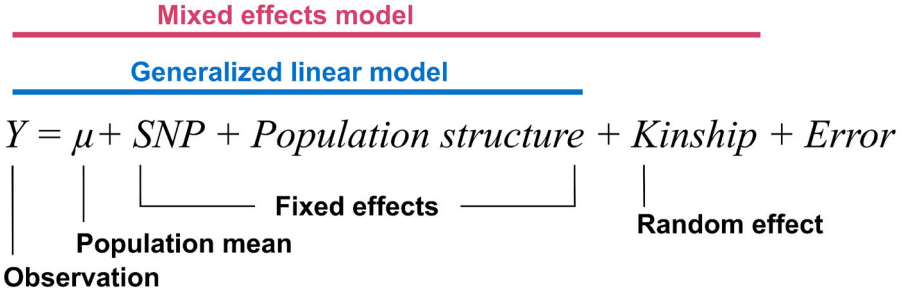

Figure 8. Generic structure of a GWAS model (adapted from Yu et al., 2006).

4.5.3. Materials

4.5.3.1. Computational resources. GWAS requires at least a computer (for small studies), up to a high-performance cluster (for large studies). Specific informatic tools such as R or other dedicated software programs (discussed below) for data manipulation, statistical analysis, and visualization are also needed.

4.5.3.2. Genotypic and phenotypic data. Genomic information can be found traditionally in the form of a genotype matrix extracted from VCF files, where different genotypes are specified as 0, 1 and 2. Alternative formats, such as allele frequencies, are also becoming more and more common and highly useful for honey bee genomic studies. A matrix linking phenotypic observations to sample identifiers will also be needed.

4.5.4. Methods

4.5.4.1. Preparation of phenotypic data. Classical GWAS methods rely on the assumption that the phenotype of interest follows a normal distribution. Therefore, depending on the trait, corrections such as log, logit, square-root, cube root or, empirical Bayes might be needed to adjust the phenotype distribution to the assumed normality. Unfortunately, as corrections need to be tested independently, no standard protocol can be described here.

We present below an example of an R script to perform empirical Bayes correction with beta distribution (Figure 9(A)). The beta distribution is often used for empirical Bayes correction as the variety of alpha and beta parameters defining the distribution allow for a large range of probability density functions representing most of the observed distribution for ratios and proportions (Figure 9(B)).

```
library(MASS) #Load library
df$pheno = df$num / df$den #estimation of the phenotype as a ratio of num (numerator) on den (denominator)
eb_fit = fitdistr (x = df$pheno[ !is.na(df$pheno) & df$pheno > 0 & df$pheno < 1 ], densfun = "beta," start = list(shape1 = 1, shape2 = 1), method = "L-BFGS-B") #fit of beta distributions to the estimated phenotype
aprior = eb_fit$estimate #alpha parameter for the beta distribution estimate for our phenotype
bprior = eb_fit$estimate #beta parameter for the beta distribution estimate for our phenotype
df$pheno_eb = (aprior + df$num) / (aprior + bprior + df$den) #corrected phenotype
```

4.5.4.2. Preparation of genotypic data. Prior to running a GWAS analysis, we recommend performing a quality control analysis on the genotypes (Wragg et al., 2021). Depending on your data, these quality controls can be based on:

- Technical and sequencing characteristics such as genotyping quality (QUAL >200 and QD >20), depth (20×), or even calling rate (>95%).
- Minor allele frequencies (MAF >0.05), or rare variants, are much harder to test accurately than common ones. Depending on the read coverage, rare variants are likely associated with genotyping errors and are often removed without further considerations.
- Linkage disequilibrium (LD), as assumed for the GWAS models, relies on the independence of the tested markers; therefore, one might want to filter on a LD threshold. Doing so will remove redundancy in the genomic information and might strengthen the signal for some markers of interest.

4.5.4.3. Performing GWAS: methods and software. Multiple software and packages are available to

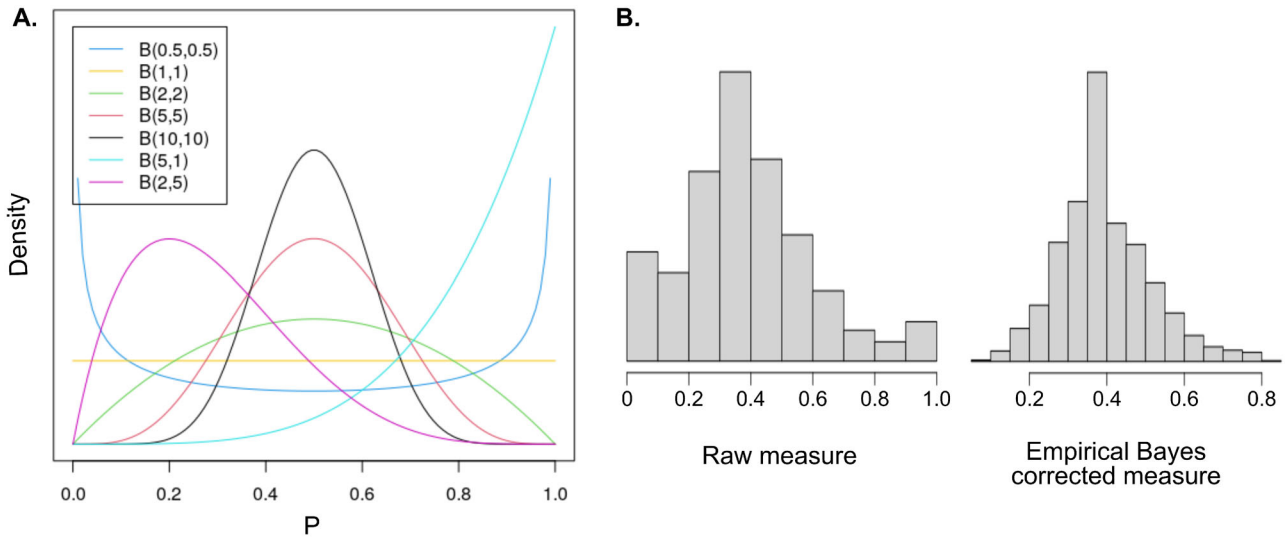

Figure 9. Correcting phenotypic parameters. (A) Example distribution of phenotypic values before and after correction using the empirical Bayes method. (B) Example of the histogram for a phenotype before and after empirical Bayes transformation. Note the improved distribution after transformation.

perform GWAS easily from genomic data (for a partial listing, see in GWAS tools platform). One standard and popular software is **PLINK** (Purcell et al., 2007). It contains a large panel of functions from data manipulation and filtering to GWAS analysis, making it a preferred tool for analyzing individual diploid data. PLINK is compatible with binary phenotype variables (case/control) as could be expected for, e.g., survival. In such a case, standard chi-square analysis on each marker individually can be used. This test gives test statistics and p-values for the difference in allele frequency between the case and control group. See Section 4.7 for more information on using **PLINK**.

However, in *Apis* honey bee colonies, traits are often measured at the group level, making pooled-sample genotype information (e.g., allele frequencies or mean genotypes) more relevant. To deal with such data types, we list some tools that have been adapted to be able to analyze pooled data. We recommend using **GEMMA** for analyzing pooled-sample data (Zhou & Stephens, 2012), which allows for mixed linear models as well as Bayesian inference. For small datasets (hundreds of samples), the **LDAK** (Speed et al., 2012) software might be more suitable for performing GWAS. Additionally, **LDAK** also offers potential GWAS computation based on gene annotation rather than genotypes. This allows a straightforward interpretation of the functional effect, since only differences within annotated genes are detected, and reduces the number of statistical tests performed. While new tools are continuously developed to perform GWAS with different settings and data types, the field still lacks dedicated methods that can deal specifically with honey bees and haplodiploidy.

4.5.4.4. Detecting signatures of selection. Although it is not solely restricted to association studies, detecting selection signals (i.e., stabilizing selection, directional selection, and diversifying selection) is of significant interest in honey bee genetic analysis to assign selection signals along the genome for specific traits (also known as phenomics). Several software options are available for this purpose, such as **FLK** and **hapFLK** (Bonhomme et al., 2010; Fariello et al., 2013). The use of such methods has helped characterize soft selective sweep in Africanized honey bees in invasive areas (Avalos et al., 2017), through a temporal sampling in native Swiss populations (Parejo et al., 2020) and in association with beekeeping practices (Wragg et al., 2016). Avalos et al. (2017) have curated customs bash and R scripts available as supplemental notes, which we recommend as a reproducible workflow using RsB for any users who want to start such analysis with honey bee samples (Tang et al., 2007).

4.5.5. Sources of variation

Population structure might need to be accounted for by adding covariates in the GWAS model. Such covariates can come from genetic background information (vectors of genetic admixture, commonly called Q vectors), principal component analysis eigenvalues for population structure, environmental effects such as the apiary, effects linked to the beekeeping practice (see Section 4.6 on population genomics). In addition, a kinship matrix (the relationship between individuals) can be included in the model. This matrix can be estimated based on pedigree records (Brascamp & Bijma, 2014), though it is often challenging for honey bees to have complete genetic records. In *Apis* honey bees, it is especially interesting to consider such a matrix, as relationships across individuals within a colony deviate from standard estimates in diploid livestock species. The software **LDAK** offers additional features. One such feature is the possibility of computing genetic relationship matrices (GRMs) from weighted polymorphisms, which is equivalent to trimming our genomic information based on linkage disequilibrium, as suggested in Section 4.5.4.2.

In case/control association studies, rigorous experimental design is advantageous, in the sense that matching case and control individuals according to environmental conditions, genetic background, etc., will authorize the researcher to ignore these variables in the analysis.

4.5.6. Quality control and data interpretation

To validate that the GWAS performed matches the model hypotheses, we recommend running multiple diagnostic analyses. The most common diagnostic is to draw the Q-Q plot for p-values of the GWAS using **R**. This method allows us to check for an agreement between expected and observed p-values. If a deviation from the expectation is observed, this is a sign of improper fit of the model to the data. This generally indicates that either the phenotype is not properly modeled or that population stratification or structure is not accounted for in the GWAS model. A strong deviation from the diagonal in Q–Q plot (Figure 10(A)) should lead to further investigation and repetition of the GWAS. Another diagnostic approach can be performed on p-values, as they are expected to follow an uniform distribution. Therefore, checking the histogram of *p*-values is a relevant diagnostic plot for a good fit of the model (Figure 10(B)). Finally, checking for unexpected correlations between SNP effects and allele frequencies

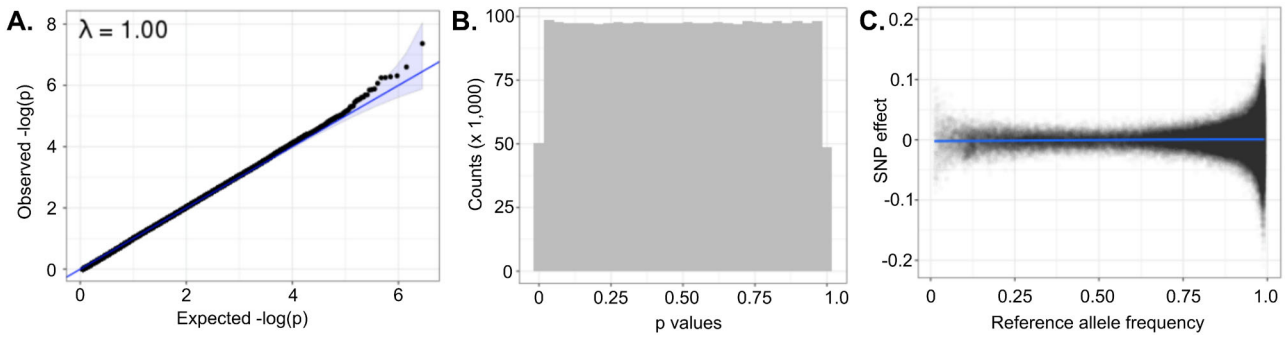

Figure 10. Example diagnostic plots for GWAS study. (A) Q-Qplot and associated inflation factor λ, (B) histogram of the *p*-values and (C) plot of SNP effects as a function of allele frequencies.

can inform on misfit of the model or of errors in our data set (Figure 10(C)).

After validation of the study based on the diagnostic plots presented above, we recommend looking into the markers' *p*-values and effects. The decision on the significant threshold cut-off for the markers associated with a trait of interest is crucial (often between 1 and 5%). Traditional corrections for multiple testing, such as Bonferroni, can be used but are often too stringent and might cause one to ignore relevant markers. For GWAS, controlling the false discovery rate (FDR) is a standard approach. Recently more methods, like ash (Stephens, 2016), are less conservative and better adapted to complex phenotypes. This method expects phenotypes to be highly polygenic with most markers having a negligible effect on the phenotype itself, as can be anticipated for complex honey bee traits. Recent developments have also made it possible to combine multiple analyses on different populations or even phenotypes using the software **Mantra** and **Mr Mega** (Morris, 2011) or **Mash** (Urbut et al., 2019).

Once diagnostic tests are completed, it is time to interpret the effects of the markers, if identified, to describe the genotype/phenotype association. In some studies, few markers may be detected with large and highly significant effects. In this case, the markers are strong candidates to explain the causal mutation underlying the trait. More often, many markers are detected, having small but significant effects. This is often the case in complex traits that tend to be highly polygenic. If markers are identified, the final step is to inspect the genomic regions of interest and identify candidate causal genes to try to interpret the biological pathways involved.

4.5.7. Applications and limitations

The pinnacle goal of GWAS is to identify genetic underpinnings of specific characteristics. The possibilities of traits to focus on are endless, but to date, association studies in honey bees have focused mostly on traits linked to beekeeping practices (Guichard et al., 2021), disease or parasite resistance (Spötter et al., 2016), and health (Wu et al., 2021). In other studies, genome-wide scans have been used to identify selection signals for population-specific features such as royal jelly production (Wragg et al., 2016). Although genetic markers for specific traits were first identified through quantitative trait loci (QTL) mapping (Evans et al. (2013) for a description of this technique), the technique has fallen out of favor as high-throughput sequencing has become more accessible.

Performing GWAS on honey bees can be challenging because the breeding system often makes the colony, which has a diverse genetic makeup, the unit of interest. Moreover, limitations arise whenever divergent and structured populations in honey bees are analyzed together without being properly accounted for in the GWAS model. For example, if bees belonging to different subspecies are analyzed together for a phenotype, the markers identified might be representatives of the different genetic types rather than linked to the phenotype of interest. This can also impact comparative genomics studies (see Section 4.4) and can lead to misguided interpretations of mechanisms underlying traits of interest.

On top of estimating the effects and significance of targeted markers associated with a trait, it is also possible to evaluate trait heritability with GWAS (Brascamp & Bijma, 2019; Guichard et al., 2021; Jourdan-Pineau et al., 2021). Such heritability measures are population-specific and might vary across studies focusing on different individuals and environments. Furthermore, genomic heritability estimates in honey bees might differ from pedigree heritability estimates, as they are not straightforward to approximate due to haplodiploidy and multiple mating (Jourdan-Pineau et al., 2021). Such heritability information is particularly useful for guiding selective breeding programs.

So far, few research projects report genome wide association study results in honey bees but have already informed on genomic regions associated with aggressive behavior (Avalos et al., 2020; Guichard et al., 2021; Sokolowski, 2020), adaptation

to mountainous environments (Everitt et al., 2023), and disease tolerance (Hassanyar et al., 2023). With the increasing availability of genomic data, due to reduction in sequencing costs and improved availability of genotyping chips, we can expect a growth in this field.

However, genome wide association studies in honey bee genomics may still face major challenges. Honey bee genomic data can have high complexity, as it may be derived from workers, drones, or pools of individuals, and there is a lack of dedicated methods to perform GWAS in haplo-diploid organisms. Furthermore, many desirable phenotypes are scored at the colony level, necessitating pooled sampling. Many of these traits (e.g., behaviors) are also complex themselves, and can be highly polymorphic, conditional, and polygenic.

Consequently, although efforts are underway (Grozinger & Zayed, 2020), it has been exceedingly challenging for GWAS results to contribute to selection decisions in the form of genomic markers with strong prediction potential for traits of interest. GWAS may be better suited to understanding the biological mechanisms underlying traits of interest for honey bee research.

4.6. Population genomics: experimental design

Whereas genome-wide association studies (GWAS) link SNPs to specific traits, population genomics identifies changes in the genetic composition of populations. To do this, population genomics analyzes hundreds to thousands, even millions, of markers (usually SNPs) across multiple individuals and populations to unravel microevolutionary patterns and processes. Due to the massive amounts of data that are typically generated by high-throughput sequencing or genotyping technologies, population genomics has evolved from classic population genetics (in which the number of markers was in the order of tens) to become a data-driven and computational science. It requires computational resources, memory, and expertise in bioinformatics due to the large-scale data generated in these studies.

In honey bees, as in many other organisms, population genomics provides important insights into the bees' demographic and adaptive history (Chen et al., 2016; Cridland et al., 2017; Fuller et al., 2015; Harpur et al., 2014; Henriques, Wallberg, et al., 2018; Nelson et al., 2017; Wallberg et al., 2014) as well as into the genomic basis of important traits, such as tolerance to diseases and parasites (Saelao et al., 2020), royal jelly production (Rizwan et al., 2020; Wragg et al., 2016), as well as defensive, scouting and recruiting behaviors (Avalos et al., 2020; Harpur et al., 2020; Southey et al., 2016).

4.6.1. Sampling strategy

When starting out a population genomics study, the first step is to design the sampling strategy. Before collecting the samples, several sampling-related aspects must be considered, including (i) sample sizes of individuals and markers, (ii) sample breadth, (iii) sampling design, (iv) sampling workers versus drones, and (v) sampling a single individual versus multiple individuals per colony. Final decisions are greatly constrained by the study's goals (e.g., estimating diversity within colonies or at the population level, inferring population structure, finding signatures of local adaptation, or landscape genomics), the complexity of the genetic patterns of the focal subspecies, and the budget available.

4.6.1.1. Sample sizes of individuals and markers.

Simulation studies have shown that population genetics inference (e.g., diversity, demographics, differentiation, or gene flow) is influenced by the sample sizes of markers and individuals (Aguirre-Liguori et al., 2020; Flesch et al., 2018; Foster et al., 2021; Landguth et al., 2012). While increasing both simultaneously produces more robust estimates, empirical and simulation studies have consistently shown that the accuracy benefits far more from increasing the number of markers than the number of individuals (Landguth et al., 2012; Nazareno et al., 2017; Willing et al., 2012).

To the best of our knowledge, there are no simulation studies on the optimal sample sizes for population genetics or genomics inquiries in honey bees. However, inferring from studies on other organisms, when the number of markers is in the order of hundreds to thousands (which is becoming commonplace in the post-genomics era) around eight individuals per population suffice for accurate estimates of diversity and differentiation (Aguirre-Liguori et al., 2020; Flesch et al., 2018; Li et al., 2020; Nazareno et al., 2017).

However, in honey bee studies, sample sizes have typically been larger. These have ranged from $n = 9$ to 87 individuals per group in population genomics studies (Avalos et al., 2020; Chen et al., 2016; Harpur et al., 2014; Henriques, Parejo, et al., 2018; Nelson et al., 2017; Wallberg et al., 2014; Wragg et al., 2016) and from $n = 12$–117 when developing SNP assays (Chapman et al., 2015; Henriques, Parejo, et al., 2018; Jones, Du, et al., 2020).

While empirical evidence from honey bee studies is lacking, the optimal number of sampled individuals will certainly vary among subspecies, depending on their distributional range or evolutionary complexity (Henriques, Browne, et al., 2018). For example, the optimal sample size for *A. m. ruttneri*, a subspecies confined to the small island nation of Malta, is

expected to be much smaller than that of A. m. mellifera, a subspecies with one of the greatest geographical distributions, or of *A. m. iberiensis*, a subspecies with a complex history involving natural hybridization. Finally, it is also important to keep in mind that unequal population samples can impact the outcomes of some analytical approaches, such as inferences on population structure (Puechmaille, 2016).

4.6.1.2. Sample breadth. In addition to the sample size, when sampling honey bees for population genomics studies, it is also important to consider the number of sampled populations, or breadth (coverage of distributional range), which can influence inferences for classic population genetics as well as landscape genetics or outlier tests (Aguirre-Liguori et al., 2020; Albert et al., 2010; Nazareno et al., 2017; Schwartz & McKelvey, 2009). In honey bees, Henriques, Parejo, et al. (2018) showed that sampling a geographically restricted area within the *A. m. iberiensis* distributional range would erroneously identify a set of SNPs with the fixation index between this and C-lineage subspecies equal to one ($F_{ST} = 1$), with an impact on the design of reduced panels of highly informative SNPs. This is because the true F_{ST} values are <1, meaning that the SNPs are not diagnostic anymore and therefore have a lower information content. When the goal of the study is to find genomic evidence of local adaptation, it is critical to ensure that populations (or individuals) are sampled across environmental gradients (e.g., latitudinal or altitudinal) and environments (e.g., arid and humid), for increased power in detecting outlier SNPs and therefore candidate genes (Manel et al., 2012).

4.6.1.3. Sampling design. The sampling design will depend on the individuals' spatial distribution and the study's goal, among other factors (Paradis, 2020). For example, if the goal is to find signatures of selection in honey bee populations, sampling should cover environmental gradients. When sampling encompasses pairs of populations that maximize the environmental differences while minimizing the evolutionary differences, there is great potential for detecting selection footprints (Delaneau & Zagury, 2012; Lotterhos & Whitlock, 2015). This approach was followed in a study of honey bee adaptation to altitude in East Africa (Wallberg et al., 2017). By sampling two pairs of populations representing mountain forests and lowland savannahs, the authors could detect strong candidates for adaptation to highland habitats in whole-genome scans.

Several sampling designs can be used by honey bee researchers. For continuously distributed populations, random and systematic sampling designs (e.g., transects or grids) have proven to be effective in simulation studies (Oyler-McCance et al., 2013). The systematic design was implemented in Iberia to unravel the genetic diversity patterns and underlying processes of *A. m. iberiensis*, *via* the establishment of three north-south transects. With this design, the authors uncovered a secondary contact zone in Iberia and found evidence for selection as another evolutionary force shaping complex genetic patterns in *A. m. iberiensis* (Chávez-Galarza et al., 2013, 2015; Henriques, Parejo, et al., 2018). If, on the other hand, the distribution is patchy, cluster sampling (sampling several groups of individuals) is more appropriate (Oyler-McCance et al., 2013).

4.6.1.4. Sampling workers versus drones. The choice of sampling workers (diploid) over drones (haploid) will depend on the questions being addressed and, thereby, on the type of analyses that will be performed (see Section 4.2 for more on general ploidy and pooling considerations). For example, Hardy–Weinberg Equilibrium (HWE) testing can only be done with individual diploid workers. However, there are analyses that do not require testing for HWE, and in this case, using haploid data can be advantageous. In addition to reducing sequencing costs, drones generate phased data, thereby circumventing the hurdles of statistical phasing. Phased genomic data offers increased power for detecting selection signatures by facilitating the employment of haplotype-based methods.

4.6.1.5. Sampling a single individual versus multiple individuals per colony. Classic population genetics analysis (e.g., HWE, diversity, differentiation, structure) requires sampling one single diploid worker per colony. This approach circumvents the problem of over-representing the queen's genotype and violating the assumption of sample independence by sampling multiple individuals. However, when the study addresses questions at the intra-colony level (e.g., patriline analysis, colony structure), sampling multiple individuals is an unavoidable requirement. The number of sampled individuals per colony can vary, depending on the questions being addressed and on budget limitations. For example, the numerous studies published in the pre-genomics era, which required patriline analysis, typically sampled tens of workers per colony (61 ± 72.6; see the review of Tarpy et al. (2004)). When multiple individuals are sampled from within a colony, they have historically been genotyped separately, but pooling is becoming increasingly popular in the post-genomics era (see Section 4.2.2).

4.7. Population genomics: filtering and summary statistics using PLINK

In this section, we will describe some tools available for analyzing SNPs (the most common type of genetic variation used in genomic studies) using a population genomics framework. We will employ a VCF tutorial file containing 3,669,288 SNPs obtained after the variant calling of 33 individuals, 26 of which are of M-lineage ancestry (18 *A. m. iberiensis* and eight *A. m. mellifera*) and seven are of C-lineage ancestry (three *A. m. carnica* and four *A. m. ligustica*).

The first step in any population genomics study is understanding and filtering the dataset by calculating summary statistics to evaluate missing data, allele frequencies, or linkage disequilibrium. To that end, several tools tailored for handling whole-genome data are publicly available, including VCFtools (Danecek et al., 2011), **PLINK** (Purcell et al., 2007), and several **R** packages such as **pegas** (Paradis, 2010) and **PopGenome** (Pfeifer et al., 2014). Here, we chose the highly versatile and efficient **PLINK** tool to illustrate the steps involved in filtering and in producing summary statistics for whole-genome datasets.

PLINK is a command-line software that runs on different operating systems such as Linux, Microsoft DOS, and macOS. Additionally, instead of using command lines, one can also use an interface through the program **gPLINK**, although this is only available for the most commonly-used commands. In this section, we will go over some of the basic commands that can be carried out with **PLINK** (version 1.9).

4.7.1. Download and installation

First, download **PLINK** 1.9 (www.cog-genomics.org/plink2/) and the tutorial dataset called pop_gen, which can be found on github.com/MaevaTecher/standard-apis-omics. All the commands will be typed at the command prompt. To check that **PLINK** is installed, type "`-./plink`" on the command prompt, and some information about the PLINK, like its version and examples of flags, will be printed.

4.7.2. Input format and conversion

PLINK can operate with various input formats, the most common are regular PLINK TXT files, *PLINK 1 binary* BIM file, and variant call format (either VCF or BCF). **PLINK 1** binary is the preferred input format because PLINK automatically converts the other formats to this one to save time and space. There are commands for converting the formats to each other. We will explain the most common files and provide the command line to convert them to others.

4.7.2.1. Variant call format (VCF). Genome sequence data are very often represented in the variant call format (VCF) or its binary counterpart (bcf). The following command can be used to convert vcf into the *PLINK 1 binary format:*

```
plink –vcf pop_gen.vcf\
- -keep-allele-order\
- -allow-extra-chr\
- -make-bed\
- -out pop_gen
```

The meanings of the flags are the following:

`- -vcf` specifies that we are using a .vcf.gz file.

`- -keep-allele-order` specifies that the reference allele will be in the 6th column (A2) and the alternate allele in the 5th column (A1). If we do not use this flag PLINK will reorder the alleles, and in A1 and A2 columns will be the minor and major alleles, respectively.

`- -allow-extra-chr` indicates that chromosome code does not start with a digit.

`- -make-bed` specifies that *PLINK1 binary files* should be generated.

`--out` specifies the output name. If one does not use `--out`, PLINK will call the file "plink," for instance "`-plink.bed.`"

If there are no errors, four files will be created: *`pop_gen.bed`*, *`pop_gen.bim`*, *`pop_gen.fam`* and *`pop_gen.log`*. The *`pop_gen.log`* file contains the command and other information that was printed to the console. The other files are explained below.

4.7.2.2. PLINK 1 binary format (.bim). To work with PLINK 1 binary format, the three following files (*`pop_gen.fam`*, *`pop_gen.bim`*, and *`pop_gen.bed`*) will be needed:

1. pop_gen.fam: Contains the sample information and has the following fields:
   - Family ID (If unknown, use the individual ID)
   - Individual ID (Alphanumeric ID to uniquely identify an individual and cannot be "0")
   - Paternal ID (If unknown, use "0")
   - Maternal ID (If unknown, use "0")
   - Sex code (Use "1" if male, "2" if female, "0" if unknown)
   - Phenotype value (Use "1" if control, "2" if case, if not used set to "0")
2. pop_gen.bim: Contains the variant information and has the following columns:
   - Chromosome (It should be a digit, otherwise one needs to use the flag - -keep-allele-order)
   - Variant identifier (Can contain any character, except spaces, tabs or "*")
   - Genetic distance in morgans or centimorgans (Can be set at "0")
   - Base-pair coordinate (Positive integers in bp)
   - Allele 1-A1 (Contains the less common allele or the alternate allele)

- Allele 2-A2 (Contains the most common allele or the reference allele)

3. pop_gen.bed: Binary biallelic genotype data.

To convert PLINK 1 binary format to VCF, one can use the following command:

```
plink –bfile pop_gene\
–recode vcf\
–out pop_gene
```

`--bfile` specifies that we are using PLINK 1 binary format.

`--recode vcf` specifies to create a VCF file as an output.

`--out` specifies the output name.

Note: To do this conversion, one needs to be sure that in PLINK 1 binary format, the reference alleles are located in the 6th column (A2).

4.7.2.3. Regular PLINK text files. The regular PLINK text files are formed by the PED and MAP files. The PED file is a white-space delimited file that stores the genotype calls. It contains six mandatory fields/columns which correspond to the `file.fam` fields. After these columns, the PED file is followed by 2*(number of variants) columns with the genotype of each locus in the same order as in the MAP file. The MAP file contains the variant information corresponding to the first four columns of `pop_gene.bim`.

The MAP file should have the same root name as the PED file and contain the same number of loci. For instance, if there are genotypes for 10 loci and 12 individuals, the MAP file will contain 10 lines and four columns, and the PED file will contain 12 lines and 26 columns (10 loci *2 + 6 mandatory columns). The loci in the PED file do not have to comply with the genomic order, but they must be in the same order as in the MAP file.

To convert PLINK 1 binary format to regular PLINK text file, one can use the following command:

```
plink --bfile pop_gene
--recode\
--out pop_gene
```

To convert a regular PLINK text file to 1 binary PLINK file, the following command can be run:

plink `--file pop_gen`\

```
--make-bed\
--out pop_gene
```

`--file pop_gen` specifies that regular PLINK text files are being used.

`--make-bed` allows the conversion to the PLINK1 binary format

If there are no errors, four files will be created: `pop_gene.bed, pop_gene.bim, pop_gene.fam,` and `pop_gene.log`.

4.7.2.4. Filtering and handling missing data. It is common for genome-wide datasets to have individuals or variants with missing data. This is particularly frequent and expected in low-coverage sequencing data. PLINK allows us to estimate the missing data per individual and loci using the flag `--missing`. A report containing the missing data per Individual ID (`pop_gen.imiss`) and variant (`pop_gen.lmiss`) can be created using the following command:

```
plink –bfile pop_gen –missing –out pop_gen
```

It is highly recommended to discard individuals or variants with high levels of missing data. The missing data thresholds are arbitrary, but values of 10% to 30% are commonly used in the literature (Henriques, Wallberg, et al., 2018; Parejo et al., 2016; Wallberg et al., 2014; Wragg et al., 2016). To remove the variants with >20% missing data `(--geno 0.2)` and individuals with >10% missing data `(--mind 0.10)`, for example, the following command can be run:

```
plink –bfile pop_gen\
--geno 0.2\
--mind 0.1\
--make-bed\
--out pop_gen_MD
```

After implementing this command, a PLINK 1 binary fileset, with the root name of "`pop_gen_MD,`" without the problematic individuals and loci, will be created. Note that PLINK first removes individuals with too much missing data, and secondly the variants.

4.7.2.5. Computing and filtering based on allele frequency. PLINK calculates allele frequencies and filters variants with a minor allele frequency (MAF) lower than an arbitrary threshold, which is commonly set at 1 or 5%. To generate a report (`pop_gene_MD.frq`) containing allele frequencies for each variant, the following command can be run:

```
plink –bfile pop_gen_MD\
--freq\
--out pop_gen_MD
```

`--freq` specifies calculating the allele frequency for each variant from this dataset.

In the report generated using this command, the first and second columns contain the chromosome and the variant ID followed by the code of the minor (3rd column) and major (4th column) variants

and the frequency of the minor allele (5th column). The last column (6th) contains the number of allele observations.

To remove the variants with MAF $<$ 5%, for instance, and create a new file set (here called `pop_gen_MD_maf005`), the following command can be run:

```
plink –bfile pop_gen_MD\
- -maf 0.05\
- -make-bed\
- -out pop_gen_MD_maf005
```

`--maf` specifies that only alleles with a minimum minor allele frequency of 5% will be kept in the dataset.

4.7.2.6. Computing differentiation indices: wright's F_{ST}.

PLINK also allows calculating the fixation index F_{ST} values for each variant using the Weir and Cockerham method (Weir & Cockerham, 1984). To do so, the flag `--fst` must be combined with the flag `--within` followed by the name of the file that specifies the sets of subpopulations to be used. In the tutorial example, the within the file, called `within_M_C_ind`, contains the Family ID in the first column, the Individual ID in the second column, and in the third column the digit 1 to identify one subpopulation (here the C-lineages subspecies) and 2 to identify the second subpopulation (here the M-lineage subspecies).

To calculate the variants' F_{ST} values between M and C-lineage, the following command can be run:

```
plink - -bfile pop_gen_MD_maf005\
- -fst –-within within_M_C_ind\
- -out pop_gen_MD_maf005
```

A report called `pop_gen_MD_maf005.fst` will be generated. This report contains six columns. The first three columns contain the chromosome number, the variant ID, and the base-pair coordinates, followed by the number of called variants or non-missing genotypes (NMISS), and the 5th column contains the F_{ST} values.

4.7.2.7. Estimating linkage disequilibrium.

Linkage disequilibrium (LD) is the non-random association of alleles at different loci. When working with whole-genome data there will be thousands of loci that are in linkage disequilibrium due to different evolutionary forces, but also, because they are physically linked. Some analyses require the use of loci that are in linkage equilibrium (not associated), and PLINK provides an easy way to do so, for example, by using the following command:

```
plink - -bfile pop_gen_MD_maf005\
   - -indep-pairwise 50 5 0.5
```

The flag `--indep-pairwise` requires three parameters: the window size, the number of SNPs to shift each step and r^2. The values of r^2 range between 0 and 1. When $r^2 = 1$, the loci provide exactly the same information. When $r^2 = 0$, the loci are in perfect equilibrium. Here, LD is calculated between each pair of SNPs using a window of 50 SNPs. If one pair of SNPs has $r^2 > 0.5$ one of them will be removed. After that, the window will be shifted to five SNPs, and the procedure is repeated until all SNP pairs are examined. Note that lower values of r^2 and bigger window sizes will remove more loci.

After running this command, two files will be created, one is named `plink.prune.in`, which contains the loci that will be retained ($r^2 < 0.5$), and the other one is named `plink.prune.out`, which contains the loci that will be discarded. These files can be used as arguments of the flags `--extract` (plink.prune.in) or `--exclude` (`-plink.prune.out`).

To extract to a new datafile the loci that are in the file `plink.prune.in`, the following command can be run:

```
plink –-bfile pop_gen_MD_maf005\
- -extract plink.prune.in\
- -make-bed\
- -out pop_gen_MD_maf005_pruneddata
```

Sometimes we do not want to discard SNPs, but we want to calculate the LD in our data. LD can be computed using the flag `- -r`$^{-2}$. However, to reduce the output size, only the pairs with $r^2 > 0.2$ and within a distance $<$1 Mb will be printed in the output by default. To modify these thresholds, the flag `--ld-window` or `--ld-window-kb` is used to specify the maximum distance between loci and the flag `--ld-window--r`$^{-2}$ is used to specify the minimum r^2. Using the following command, the r^2 values will be obtained for all pairs of loci with $r^2 > 0.5$ in a window of 50 Kb. The output called `pop_gen_MD_maf005_pruneddata.ld` will be created, and the r^2 values will be placed in the 7th column.

```
plink –bfile pop_gen_MD_maf005_pruneddata\
- -r2\
- -ld-window-kb 50\
- -ld-window-r2 0.5\
- -out pop_gen_MD_maf005_pruneddata
```

`--r`$^{-2}$ specifies to compute the linkage disequilibrium.

`--ld-window-kb` specifies the maximum distance between two loci in kilobases.

`--ld-window--r`$^{-2}$ specifies the minimum r^2 to be printed out.

PLINK also allows haplotype blocks to be detected in the genome using the flag `--blocks`. If arguments are not provided to this flag, the default options will be implemented. One of them is related to the maximum block size, which is by default 20 kb. To cancel this default flag, the

argument `no-small-max-span` can be used. The flag `--blocks` without arguments assumes that in the dataset, there are phenotypes in the 6th column of the PED or FAM files. If there is no phenotype data, the argument `no-pheno-req` is required. By default, the pairwise LD is only calculated for variants within 200 kb. If needed, this parameter can be changed using the flag `--blocks-max-kb`. An example of a command to calculate blocks along the honey bee genome can be:

```
plink –bfile pop_gen_MD_maf005\
–blocks no-pheno-req no-small-max-span\
–blocks-max-kb 5000\
–out pop_gen_MD_maf005
```

Two files called `pop_gen_MD_maf005.block` and `plink.blocks.det` will be created. Each line of these files represents a block. The file `pop_gen_MD_maf005.block` contains the variant IDs found in each block whereas `plink.blocks.det` contains their positions. If there are 10 variant IDs or positions in one line, it means that the block contains 10 SNPs.

4.8. Population genomics: inferring population structure using ADMIXTURE

To study the genetic differences among individuals, a population structure analysis can be performed. Many analytical tools have different assumptions. The principal component analysis (PCA) makes the least number of assumptions and is suited for a preliminary data view. There are also clustering methods that estimate the proportion of ancestral populations in each individual. The software **STRUCTURE** (Pritchard et al., 2000) is the classical tool for structure analysis, however, it is not suited to large datasets. Examples of clustering methods that can handle large datasets are **fastStructure** (Raj et al., 2014), FRAPPE (Tang et al., 2005), and the widely used **ADMIXTURE** (Alexander et al., 2009). The results obtained with population structure analysis are necessary before performing more sophisticated analyses. However, the results should not be over-interpreted because different demographic scenarios may produce similar results. Here, we will show a protocol for running **ADMIXTURE**.

ADMIXTURE (version 1.3) uses a maximum likelihood approach to estimate individuals' ancestry from multilocus SNP data. It requires unrelated individuals and does not explicitly consider LD, so we advise avoiding strong LD in the dataset.

4.8.1. Download and installation

Download and install PLINK (see Section 4.7), the online tool CLUMPAK (Clustering Markov Packager Across K (Kopelman et al., 2015), and ADMIXTURE (version 1.3).

4.8.2. Input files

ADMIXTURE accepts **PLINK** and **EIGENSTRAT** files as inputs. Here, the binary 1 **PLINK** format will be used. Because of the **ADMIXTURE** assumptions, the file named `pop gen_MD_maf005_pruneddata` generated in Section 4.7.2.7 will be used. Note: In case the relatedness of the individuals under study is unknown, plink2 and the following command line can be run:

```
plink2 --bfile out_dataset\
--king-cutoff 0.177\
--make-bed\
--out relpruned_data
```

4.8.3. Methods

ADMIXTURE estimates individuals' ancestry considering a specific number of source populations commonly denoted by K. **ADMIXTURE** can be run for a predefined K when the number of source populations is known. Here, K can be set at two because the individuals used in the tutorial are of M- and C-lineage ancestry). To run **ADMIXTURE** for $K = 2$, the following command can be run:

```
./admixture pop_gen_MD_maf005_pruneddata.bed 2
```

Two files will be produced: the `pop_gen_MD_maf005_pruneddata.2.Q`, which contains the proportions of each cluster for each individual, and the `pop_gen_MD_maf005_pruneddata.2.P`, which contains the allele frequency of the source population for each SNP. However, commonly, the best K is unknown and should therefore be calculated. To do that, we should run ADMIXTURE for different Ks and use the flag `--cv` to enable the calculation of cross-validation errors. A way of doing that is through a bash command-line code, which we can call, for instance, `CV.sh`.

Step 1. Create and open the file `CV.sh`:

```
nano CV.sh
# or you can use any text editor such Vim
vi CV.sh
```

Step 2. Write a script to run **ADMIXTURE** for different Ks (here from 1 to 8) and with the flag `--cv`:

```
#!/bin/bash -l
for K in {1.8}; do
./admixture --cv pop_gen_MD_maf005_pruneddata.bed ${K}
done
```

Step 3. Run the script created in step 2.

```
chmod + x CV.sh
./CV.sh > admixture.out
```

Step 4. In the file `admixture.out` we have everything that was printed on the console, but we want to have just the CV values. For that, we can use the following command:

```
grep "CV" admixture.out | awk "{print $3,$4}" | cut -c 4,7-
  20 > admixture.cv.error
```

We can now use the cross-validation error values in the file `admixture.cv.error` to construct a plot (Figure 11(A)). The smaller the cross-validation error value, the better the prediction. In the tutorial example, $K = 2$ is the best, suggesting the existence of two source populations. However, choosing the best K can be a difficult endeavor, and another approach is to run different Ks and use the one that makes the most biological sense.

In addition, to run different Ks we should also keep in mind that ADMIXTURE can produce different outcomes for replicate runs. A good practice is to perform several runs for each K with different random seeds (or starting points), which can be done using the flag `-s time`.

In the tutorial example, ADMIXTURE will be executed 20 times for each K. To save all the runs, we must change the name of the outputs, a task that is performed by the command `mv`. This command will change the names by appending the K and the run number. For example, consider `pop_gen_MD_maf005_pruneddata.4.10.Q` is the 10th result for $K = 4$.

The script below should be written in a file called, for instance, `running_admixture.sh`. To run this, the same steps described for `-CV.sh` should be performed.

```
#!/bin/bash -l
for RUN in {1.20}; do
for K in {1.8}; do
./admixture -s time pop_gen_MD_maf005_pruneddata.bed ${K}
mv pop_gen_MD_maf005_pruneddata.${K}.P
pop_gen_MD_maf005_pruneddata.${K}.${RUN}.P
mv pop_gen_MD_maf005_pruneddata.${K}.Q
pop_gen_MD_maf005_pruneddata.${K}.${RUN}.Q
done
done
```

The last step is generating a plot showing the ancestry of each individual. An easy way of doing that is by using the online tool called **CLUMPAK** (Kopelman et al., 2015). **CLUMPAK** uses Distruct (Rosenberg, 2004) to generate the plot and **CLUMPP** (Jakobsson & Rosenberg, 2007) to align the multiple **ADMIXTURE** runs. When different solutions are found for the same K, **CLUMPP** will group the most similar solutions into "modes." To run **CLUMPAK**, one just needs to use the Q-matrices of the runs. A folder containing all the Q files produced by ADMIXTURE can be created. In this example, such a folder can be called Q_values and then zipped. Next, on the "Main Pipeline" page, one just needs to choose the option **ADMIXTURE** and upload the zip folder.

The graphic output is shown on the main page for each K's major modes (Figure 11(B)). Each individual is represented by a thin vertical line, and each color represents an inferred ancestral population. The ancestry proportion corresponds to the length of the color in the vertical bar. Only one mode was found for $K = 1$, 2 and 3. However, for $K = 4$, the major mode is represented by 18 out of 20 runs, meaning that these 18 have similar solutions. Cross-validation testing determined that $K = 2$ was the most likely outcome.

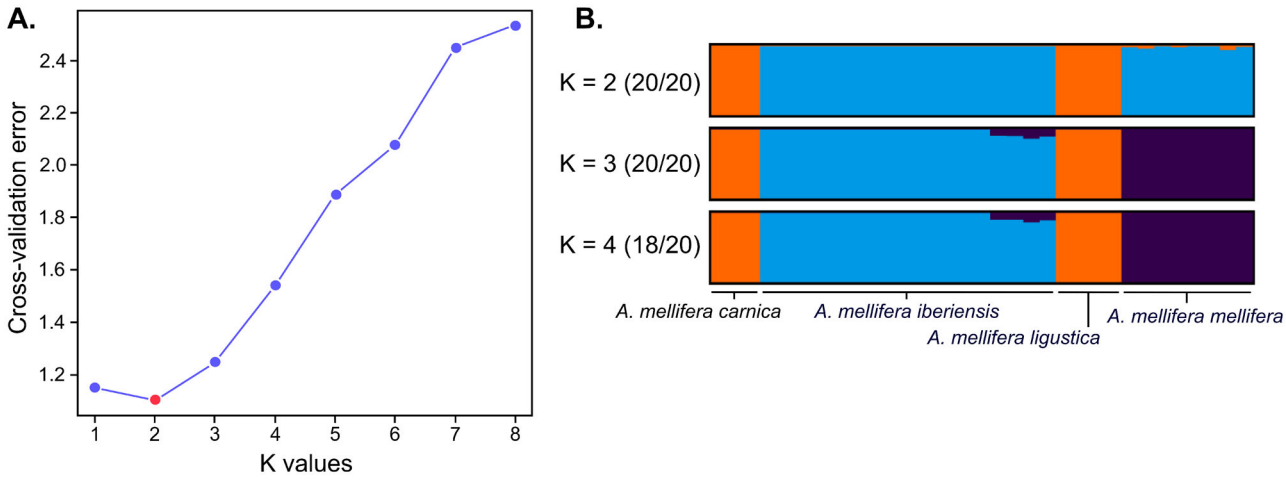

Figure 11. Example of population structure graphical output using ADMIXTURE in four European honey bee subspecies. (A) Plot of cross-validation error obtained by ADMIXTURE from $K = 1$ to $K = 6$. The lowest error was found for $K = 2$ and is marked in red. (B) Graphical representation of ADMIXTURE results obtained by CLUMPAK. Here, the populations of *A. m. carnica* and *A. m. ligustica* are represented in orange, and *A. m. iberiensis* and *A. m. mellifera* are in blue. Of note is that some *A. m. mellifera* individuals have a small percentage of orange color, suggesting a history of introgression.

4.9. Landscape genomics: an example using LFMM

Landscape genomics, which addresses how local environmental conditions influence the distribution of genetic variation (Manel et al., 2003), is an increasingly important field. To identify associations between environmental variables and genomic variation, various approaches known as genotype-environment association (GEA) or environmental association analysis (EAA) have been developed. Examples of GEA-based software are **Bayenv2** (Günther & Coop, 2013), **LFMM** (Frichot et al., 2013), and **Samβada** (Stucki et al., 2017).

When compared to F_{ST}-based methods, GEA methods can potentially identify selective pressures driving local adaptation. Besides, the combined analysis of genetic data and environmental data increases the power to detect loci that may be under selection (De Mita et al., 2013; de Villemereuil et al., 2014; Rellstab et al., 2015). However, each GEA method has advantages and disadvantages (see Rellstab et al., 2015 for further details), and combining different approaches helps reduce uncertainty of the results. Here, we will describe how to use the latent factor mixed model approach (**LFMM**).

LFMM is available in a command line version (1.5), and it can be downloaded from https://bcm-uga.github.io/lfmm/index.html. The latest version (2.0) is also available in an R package called lfmm. This tool models the effect of population structure using latent factors; therefore, population structure should be previously investigated. Here, the results obtained in the Section 4.8 will be integrated in this example.

4.9.1. Materials

To run **LFMM**, two datasets will be required:

- Genomic dataset: The PLINK 1 binary file obtained in Section 4.7.2.5, called `pop_gen_MD_maf005`.
- Environmental variables dataset: A file called `ex_env.csv`, containing 36 variables for each individual.

To have the environmental variable dataset the geographical coordinates of each apiary are needed. These coordinates are then used to obtain environmental variables from publicly available databases such as WorldClim, Climatic Research Unit, and OPENEI. A function within the **R.SamBada** package, (Duruz et al., 2019) called `createEnv`, can help in this process (**R.SamBada** documentation can be found at https://cran.r-project.org/web/packages/R.SamBada/R.SamBada.pdf).

4.9.2. Methods

To prevent problems caused by non-independency of environmental variables, first, it is important to determine if they are correlated with each other. To do that, the R packages ade4 (Thioulouse et al., 2018) for PCA computation, and factoextra (Kassambara & Mundt, 2017) for PCA visualization, will be needed. The following steps are pursued in R code:

Step 1. Install and open the R libraries:

```
install.packages("ade4") #PCA computation
install.packages("factoextra") #PCA visualization
library(ade4)
library(factoextra)
```

Step 2. Read the file with environmental variables:

```
ex_env <- read.table("ex_env.csv",
header = TRUE, #The file contains a header
row.names = 1, #Row names are in the first column
sep = ";")
```

Step 3. Perform a PCA with the environmental variables:

```
ex_env.pca<- dudi.pca(ex_env,
scannf = FALSE, #Hide screen plot
nf = 5) #Number of components kept in the results
```

Step 4. Make a correlation circle

```
s.corcircle(ex_env.pca$co)
```

A correlation circle similar to that of Figure 12(A) will be obtained. Here, the direction and the length of arrows indicate the correlation between the variables as well as between variables and principal components. For instance, in Figure 12(A), the variables AUTU_TMP and TMN_ANNUAL have the same direction, the same length and are very close to each other, which indicates that they are highly correlated. Indeed, the correlation value (r^2, calculated in the next step) is 0.99. On the other hand, the variables pointing to opposite sides of the graphs, for instance SUM_TMX and AUTO_CLD, are negatively correlated ($r^2 = -0.80$).

Step 5. Calculate a correlation matrix using the following commands:

```
cor_matrix = cor(ex_env, method="pearson") #To calculate the
  correlation values
write.csv(cor_matrix, file = "cor_matrix.csv") #To save the matrix in a
  file
```

Following that, a file called cor_matrix.csv will be created. The values with $|r| > 0.80$ should be deleted from the environmental dataset. After this, the same procedure described above (Steps 1–4) should be performed again to obtain a diagram similar to

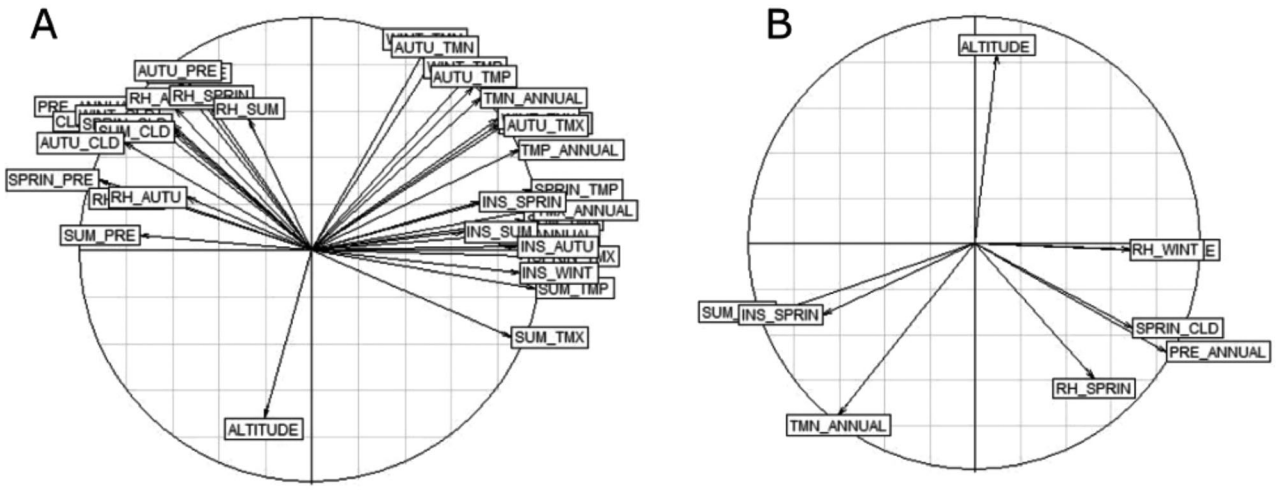

Figure 12. Correlation circles with (A) 36 variables and (B) 8 variables, after removing the highly correlated variables ($|r| > 0.80$).

Figure 12(B) to check that the represented variables (eight in total) are independent.

Now that the bioclimatic dataset is prepared, the genetic dataset needs to be handled. LFMM accepts its own format (.lfmm), and the software has a function to convert PED format to .lfmm.

Step 6. Convert the `pop_gen_MD_maf005` binary PLINK 1 binary file obtained in Section 4.7.2.5 to a regular PLINK text file.

```
plink –bfile pop_gen_MD_maf005 –keep-allele-order –recode –out
```

Step 7. Convert the PED file to .lfmm using the **R** function of **LEA**, `ped2lfmm`. The tutorial dataset contains 33 individuals and 2,365,431 SNPs. In order to efficiently upload this data into R, the **data.table** R package will be used.

```
# To install LEA
if (!requireNamespace("BiocManager," quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("LEA")

#To download the package data.table
install.packages("data.table")

#Loading the package lea
library(LEA)

#Loading the package data.table
library("data.table")

#Convert ped format to lfmm
output <- ped2lfmm("pop_gen_MD_maf005.lfmm")

#Loading the ped file to R
geno <- fread("pop_gen_MD_maf005.ped," showProgress = FALSE)
```

Note that, when the format PED is converted to the lfmm format, the names of the SNPs that were in the MAP file will be lost.

Step 8. Upload the eight independent environmental variables, but without the individuals' names that are in the first column.

```
#Load the reduced environmental file
env <- fread("ex_env_red.csv")

#Select from column 2 to 9
env = env[,2:9]
```

Step 9. Perform the association analysis between the environmental variables and the genotypes. To that end, we need to run a function (`lasso_lfmm`) to estimate the latent factor mixed model parameters. The results of this function will be used by the function `lfmm_test` to calculate the significance of the associations.

```
#To download the package lfmm
install.packages("lfmm")

#To load the package lfmm
library("lfmm")

lfmmlasso <- lfmm_lasso(Y = geno,#The genotypes matrix
X = env, #The environmental matrix
K = 2, #The best K calculated previously
nozero.prop = 0.01)

pv <- lfmm_test(Y = geno, X = env,
lfmm = lfmmlasso, #The results obtained previously
calibrate = "gif")
```

Step 10. The variable `pv` is a list and contains different types of information, such as the p values and the calibrated p values for all eight environmental variables. If one wants to see the calibrated p values of the first loci the following command should be typed:

```
head(pv$calibrated.pvalue)
```

Step 11. This command will retrieve the first six calibrated p values for each one of the eight environmental variables. Each environmental variable is in a specific column. For instance, altitude is in the first column and SUM_PRE in the third column. If we

want to see only the *p*-values of SUM_PRE, we just need to specify that this variable is in the third column, using the following command:

```
head(pv$calibrated.pvalue[,3])
```

Step 12. Select one environmental variable to continue with the procedure. The data obtained for SUM_PRE (third column) will be used. Here, the calibrated p values will be saved in the variable `pvalues_pre_ver.`

```
pvalues_pre_ver <- pv$calibrated.pvalue[,3] #SUM_PRE is the third
  column
```

Step 13. Make a Manhattan plot coloring the SNPs with *p* value <0.0001, which corresponds to -log10(*p* value) >4, in red and the others in grey.

```
#The name of the image to be created
jpeg(filename="Manhattan.jpeg")

#Plot the – log10 of the calibrated p-values
plot(–log10(pvalues_pre_ver), pch = 19, #Each p-value is represented
  by a circle
cex = 0.2, #To specify the size of the circle
xlab = "SNP," ylab = "–Log P,"#To label X and Y axis
col = ifelse((–log10(pvalues_pre_ver)) > 4,"red,""grey")) #Values >4 in
  red and the others in grey

#To save the image in the working directory
dev.off()
```

Step 14. Count the SNPs significantly associated with SUM_PRE (*p* value <0.0001) and save the positions of the SNPs significantly associated with SUM_PRE in a variable called `p -value_00001_pos`. In the tutorial example, there are a total of 1,215 SNPs with p value <0.0001.

```
length(pvalues_pre_ver[pvalues_pre_ver < 0.0001])
pvalue_00001_pos = which(pvalues_pre_ver < 0.0001)
```

Step 15. Create a file in the directory with the information contained in `-p-value_00001_pos`.

```
write.table(pvalue_00001_pos,"pvalues_
  00001.csv,"row.names = FALSE, col.names = FALSE)
```

Step 16. Get the genomic coordinates from the file positions of the outlier SNPs. Use the *pop_gen_MD_maf005.map* coordinates and retrieve only the positions indicated in *pvalues_00001.csv*.

To that end, the following steps can be performed:

1. Create a file called *index2snp.sh*.

```
nano index2snp.sh
```

2. Type the following code in the file and close it.

```
#!/bin/sh
cat pvalues_00001.csv | while read line; do
sed -n ${line}p pop_gen_MD_maf005.map >> pvalues_00001_snp_
  name
done
```

3. Allow executing of the file.

```
chmod + x index2snp.sh
```

4. Run the file.

```
./index2snp.sh
```

Step 17. Annotate the outlier SNPs. To that end, the tool SnpEff (Cingolani et al., 2012) can be used in the Galaxy server. SnpEff accepts as input a VCF file. Thus, the first step is to build a VCF only with the outlier SNPs, which can be done in **PLINK**. To reduce the size of the VCF file, only one individual will be included in the tutorial example. The following steps are required to annotate the outlier SNP:

1. Create a file with the Family and Individual ID that will be retained. In the example by writing "car ID1" in the file *one_ind*.

```
nano one_ind
car ID1
```

2. Create a file with the SNPs list to be retained.

```
awk "{print $2}" pvalues_00001_snp_name > outlier_snp
```

3. Create a VCF file.

```
plink --bfile pop_gen_MD_maf005\
--keep one_ind `# list of individuals to keep `\
--extract outlier_snp `# list of SNPs to keep`\
--recode vcf –out outlier_snps
```

4. Go to Galaxy and select the tool "SnpEff download" and write "*Apis_mellifera*" in the box "Select the annotation database you want to download."

5. While on the Galaxy platform, go to the "SnpEff eff" tool and upload the *outlier_snps.vcf* file. In "Genome source," select "Downloaded snpEff database in your history" and in "Genome data," select the database downloaded in the previous step.

After this step, Galaxy will provide a report with statistics (in input HTML stats file) such as the number of missense variants and the number of SNPs per chromosome. A second file shows all the genes and regions in which SNPs are located. Similar to many other selection studies, most selective SNPs detected in the dataset tested herein fall outside of exons (609 intergenic and 573 intronic). There were only 25

exonic SNPs, two of which appear to be missense or non-synonymous, and these were located in candidate genes GB42150 and GB46140.

While genome-wide scans provide a powerful way of highlighting candidate genes for selection, causal inferences about the molecular basis of adaptation can be obtained from functional and expression studies and also from *in silico* protein modeling (Henriques, Wallberg, et al., 2018). Nonetheless, before employing such approaches, cross-validation of detected outlier SNPs should be always sought by other conceptually different methods.

4.10. Applying population genomics to conservation: reduced SNP analysis

Whole genomes provide important insights into the processes that shape diversity patterns (Chen et al., 2016; Cridland et al., 2017; Fuller et al., 2015; Harpur et al., 2014; Henriques, Wallberg, et al., 2018; Nelson et al., 2017; Parejo et al., 2017; 2020; Wallberg et al., 2014) and are offering unprecedented power for delving into fundamental apicultural questions (e.g., tolerance and resistance to the ectoparasite *Varroa destructor*, introgression, etc.) (Harpur et al., 2019, 2020; Parejo et al., 2016; Saelao et al., 2020)), with potential implications for sustainable honey bee management. However, the adoption of whole genome approaches is hindered by the requirement for sophisticated laboratory and computational resources, as well as advanced bioinformatics expertise. Many molecular biology laboratories, conservation centers, and breeding facilities lack access to these resources, limiting the value of genomic data in these contexts. To address this challenge, experts can leverage whole genome data to develop reduced SNP-based tools, and these can be more easily employed by the broader honey bee community.

Panels containing a reduced number (<160) of highly informative markers have been designed from genome-wide SNPs (Muñoz et al., 2017) or from whole-genome data (Chapman et al., 2015; Henriques, Parejo, et al., 2018) to address different goals such as (i) identifying Africanized honey bees (Chapman et al., 2015), (ii) estimating C-lineage introgression into the M-lineage *A. m. mellifera* and *A. m. iberiensis* subspecies (Henriques, Parejo, et al., 2018; Henriques, Wallberg, et al., 2018; Muñoz et al., 2017) or (iii) monitoring diversity in immune genes (Henriques et al., 2021). All these reduced SNP panels have been tailored for genotyping in the MassARRAY MALDI-TOF platform, a cost-effective technology for a relatively small number of markers and a large number of samples. For example, in a single 384 SpectroCHIP array, it is possible to genotype 384 samples with one SNP-plex (containing a maximum of 40 SNPs), 192 samples with two SNP-plexes, 128 samples with three SNP-plexes, 96 samples with four SNP-plexes, or other possible combinations of markers and samples to a maximum of 384, all for the same price. Additional technologies, such as Fluidigm microfluidic array, SNaPshot, and capture-based target enrichment methods, are also appealing for the development of genotyping tools because they also allow screening dozens to hundreds of SNPs in a cost-effective manner (Daca-Roszak et al., 2016; Emerman et al., 2017; von Thaden et al., 2020).

When the number of selected SNPs is in the order of thousands to tens of thousands, other technologies such as Affymetrix or Illumina Infinium are better suited for genotyping. These technologies have been used for genotyping highly dense SNP panels with varying purposes, including genome-wide association screening for hygienic behavior (∼44,000 SNPs) (Spötter et al., 2016), identification of honey bee subspecies (∼4,000 SNPs) (Momeni et al., 2021), or genomic selection (∼100,000 SNPs) (Jones, Du, et al., 2020).

When developing a panel, the first step is to establish a clear goal that will guide the choice of SNPs to be included. For instance, if one wants to create a panel for screening populations for local adaptation, a selection analysis such as LFMM should be performed. However, if the goal is to identify introgressed populations, an admixture analysis with the focal subspecies should be performed. To develop a reliable molecular tool, it is important to start out with a powerful discovery panel, which should include a reasonable number of individuals that capture, as much as possible, the entire population diversity. Using *A. m. iberiensis* as the model organism (Henriques, Parejo, et al., 2018) demonstrated that, when employing a sample size <10 in the panel design and a sample breadth representing only a fraction of a population's genetic diversity, a bias is introduced in the informativeness of the markers. This finding implies that to develop a reliable reduced SNP panel, one must first understand the diversity patterns of the target population.

This section shows all the steps involved in developing a reduced panel tailored to distinguish M- from C-lineage subspecies. In this tutorial, the panel contains only 40 highly informative SNPs, the maximum plex size of the MassARRAY MALDI-TOF technology.

4.10.1. Materials

This section requires PLINK (see Section 4.7), CLUMPAK, ADMIXTURE (version 1.3; see Section 4.8), snpEff from Galaxy (see Section 4.9), and the R package chromoMap. The dataset used herein will be

`pop_gen_MD_maf005`, which was created in Section 4.7.2.5.

4.10.2. Methods

Step 1. Divide the dataset into training and holdout subsets, following Anderson's method (Anderson, 2010). This is a critical step because if the informative SNPs are chosen and validated on the same individuals, an upward bias will be introduced. The training subset should contain 75% of randomly chosen individuals, which will be used to select the most informative SNPs. The remaining 25% of the individuals will make the holdout subset, which will be used to validate the SNP panel.

1. Create a file with the holdout subset. The individuals should be chosen at random, but to follow the tutorial we provide the file called `holdout`.
2. Create the holdout subset containing only the individuals listed in the `holdout` file:

```
plink –bfile pop_gen_MD_maf005\
- -keep holdout `#list of individuals to keep` \
- -make-bed\
- -out pop_gen_MD_maf005_holdout
```

3. Create the training subset excluding the individuals listed in the holdout file:

```
plink - -bfile pop_gen_MD_maf005\
- -remove holdout `#list of individuals to removè \
- -make-bed\
- -out pop_gen_MD_maf005_training
```

Step 2. Select the most informative SNPs from the training dataset. The main goal of the panel designed herein is to distinguish M- from C-lineage subspecies. Given that these two lineages are very divergent, it is expected to find many fixed SNPs (F_{ST}= 1), which are the most informative ones.

1. Calculate the F_{ST} values between C- and M- lineage for each SNP in the training subset (see Section 4.7.2.6. for further details).

```
plink - -bfile pop_gen_MD_maf005_training\
- -fst - -within within_M_C_ind_training\
- -out pop_gen_MD_maf005_training
```

2. Select the SNPs with F_{ST} = 1. First, the `-awk` command will select the lines that have a value equal to one in the fifth column and, from these values, only the second column will be printed in the file `pop_gen_MD_maf005_training.fst_1`.

```
awk "{if($5 == 1){print}}"
pop_gen_MD_maf005_training.fst |\
awk " {print $2} " >
pop_gen_MD_maf005_training.fst_1
```

Step 3. Apply filters to narrow down the number of discovered SNPs. In this tutorial example, a total of 416,123 SNPs exhibit an F_{ST} = 1. To narrow down this massive number, filters related to the panel goal must be applied. Here, only SNPs classified as having a high impact by **snpEff** will be retained. Other filtering criteria include linkage (SNPs that are physically linked may contain redundant information and can therefore be eliminated) and coverage of the 16 honey bee chromosomes. To investigate how the SNPs are distributed across the chromosomes, the R package chromoMap will be used. This package requires two files: the `chromosome_file.txt`, containing the chromosome coordinates, and the `-high_impact.txt`, containing the genomic coordinates of the SNPs.

1. Convert the `pop_gen_MD_maf005_training` PLINK 1 binary file to VCF format:

```
plink - -bfile pop_gen_MD_maf005\
- -extract pop_gen_MD_maf005_training.fst_1\
- -recode vcf –out FST1snps
```

2. Using the methodology described in Step 17 of the LFMM section (Section 4.9), annotate the file FST1snps.vcf with snpEff.
3. Download the annotated VCF file from Galaxy, here named `Galaxy21.vcf`.
4. Create the annotation file to be used by chromoMap. The command `grep` will search and retrieve the lines that contain the string "HIGH." From these lines, the command `awk` will print the third column (SNP name) followed by the first column (chromosome name) and the second one (genomic position) two times.

```
grep "HIGH" Galaxy21.vcf |\
awk " {print $3"\t" $1"\t" $2"\t"$2} " > high_impact.txt
```

5. Visualize the SNPs distribution using chromoMap in R.

```
install.packages("chromoMap")
library(chromoMap)
chromoMap("chromosome_file.txt," "high_impact.txt," ploidy = 1)
```

6. As depicted in Figure 13, some SNPs are located in close proximity to each other. Use PLINK to remove one of these linked SNPs by running the following commands:
   a. Create a SNP list with the SNPs annotated with a high impact

```
grep "HIGH" Galaxy21.vcf |\
awk " {print $3}" > high_impact_list
```

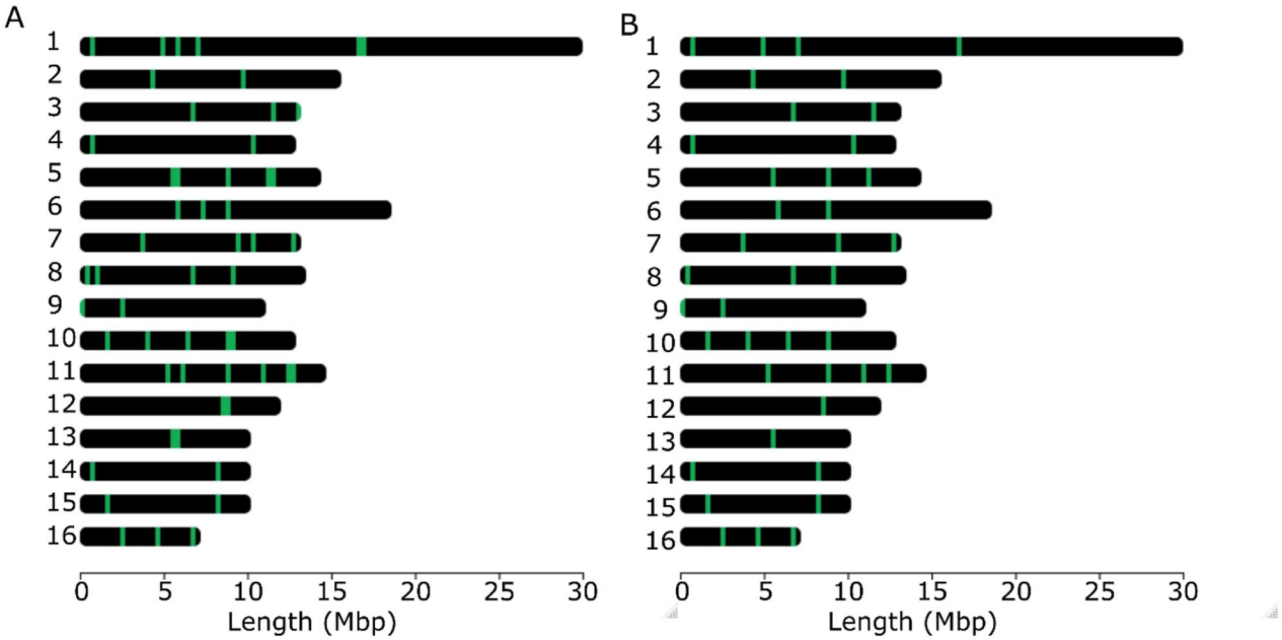

Figure 13. Chromosome map showing the positions (green line) of the highly informative SNPs (A) before and (B) after linkage disequilibrium pruning.

b. Create a PLINK 1 binary file containing only the high impact SNPs.

```
plink –bfile pop_gen_MD_maf005\
–extract high_impact_list\
- -make-bed\
- -out pop_gen_MD_maf005_high_impact_list
```

c. Create a file excluding SNPs that are in the same genomic region. To have a final list of 40 SNPs, only loci that are at least 1,500 Kb apart will be selected (using the flag `--bp-space`).

```
plink - -bfile pop_gen_MD_maf005_high_impact_list\
- -bp-space 1500000\
- -make-bed\
- -write-snplist - -out pop_gen_MD_maf005_high_impact_list_prunned
```

Step 4. Assay validation. Using the holdout subset, it is possible to assess whether the admixture proportions inferred from whole-genome data are similar to those obtained with the reduced SNP panel.

1. Run **ADMIXTURE** following the steps detailed in Section 4.8 using all the SNPs in the holdout subset (file *pop_gen_MD_maf005_holdout*)
2. Run **ADMIXTURE** for the holdout subset using only the highly informative SNPs (file *pop_gen_MD_maf005_high_impact_list_pruneddata*)
3. Compare the results obtained in (1) and (2) by calculating, for instance, Pearson's correlation (r).

In the tutorial example, all tested individuals were revealed to be pure (with no signs of introgression) and $r = 1$. If one wants to develop a reduced SNP panel for estimating introgression proportions, the panel should also be validated in admixed individuals.

5. Epigenomics

5.1. Introduction

Genetic components of genes are the fundamental basis of molecular mechanisms for social evolution (Rehan & Toth, 2015). Gene expression and regulation are precisely programmed not only by the sequence of the DNA, but also by epigenetic modifications occurring on the genomic DNA (Tirado-Magallanes et al., 2017; Yong et al., 2016), on messenger RNA (mRNA) (Peer et al., 2017; Wang, Xiao, et al., 2021), proteins that interact with DNA (Shirvaliloo, 2022), chromatin accessibility (Liu et al., 2019) transcription factor binding motifs (Hu et al., 2010), and other non-coding RNA (Benayoun et al., 2015; Ruffo et al., 2023). These factors collectively contribute to the intricate programming of gene expression and control.

Honey bees are known to be among the arthropods embodying remarkably the phenomenon of phenotypic plasticity, particularly regarding development, reproductive abilities, and behavior (Corona et al., 2016; Pfennig et al., 2010). In a colony, there are not only different castes (such as queens and workers) but also different divisions within workers (such as nurses and foragers), all of which can arise from the same genome. Many of the differences between these phenotypes responding to change in environmental cues (e.g., nutrition, pheromones,

colony size and stress) are driven by epigenetic changes (Figure 14(A,B)) which are molecular modifications that regulate genes without changing the actual DNA sequence (Feinberg, 2007; Haig, 2004; Jirtle & Skinner, 2007). DNA methylation (Glastad et al., 2014; Herb et al., 2012; Kucharski et al., 2008; Oldroyd & Yagound, 2021; Shi et al., 2011), RNA methylation (Wang, Xiao, et al., 2021), histone modifications (Glastad et al., 2019), microRNAs (Ashby et al., 2016; Behura & Whitfield, 2010; Shi et al., 2012), other non-coding RNAs (Glastad et al., 2019; Tadano et al., 2009), or organization and material state of chromatin (Wojciechowski et al., 2018) are all examples of epigenetic mechanisms. However, DNA methylation and histone modifications are among the most commonly studied, and new attention is being given to RNA methylation. Therefore, this chapter focuses on these three types of epigenetic modifications.

Honey bees have a DNA methylation system with two key enzymes: DNA-methyltransferase 1 and 3 (DNMT1 and DNMT3) (Wang et al., 2006). The most common covalent modification of DNA is the methylation on the position 5 of the cytosine base (5-methylcytosine, or 5mC), which often appears with guanine (CpG) (Li-Byarlay, 2016; Maleszka, 2008; Ruden et al., 2015; Wang & Li-Byarlay, 2015; Yan et al., 2014). Compared to vertebrate species, the honey bee genome has lower coverage of CpG DNA methylation (Bewick et al., 2017; Schmitz et al., 2019).

Environmental factors, such as changes in nutrition, biotic stress and other external stimuli (e.g., resource availability), can affect honey bee caste determination, development, and behavior (Drewell et al., 2014; Foret et al., 2012; Galbraith et al., 2015; Herb et al., 2012; Kucharski et al., 2008; Li-Byarlay et al., 2020; Lyko et al., 2010; Remnant et al., 2016). These changes are often mediated by epigenetics. Epigenetics is, therefore, a bridge linking genetics and the environment, contributing to the complex biology of honey bees (Wang & Li-Byarlay, 2015). However, the parent-specific gene expression, as associated with the kinship theory of intragenomic conflict, did not involve DNA methylation (Wu, Galbraith, et al., 2020).

5.2. DNA methylation

5.2.1. Bisulfite-seq

Bisulfite sequencing (BS-seq) is a method used to study DNA methylation at the single-nucleotide resolution. The method involves treating DNA with sodium bisulfite, which converts unmethylated cytosine residues to uracil, while leaving methylated cytosine (5mC) residues unchanged. Consequently, after bisulfite treatment, only the methylated cytosines are retained. The treated DNA is then sequenced, and the resulting data can be analyzed to determine the methylation status of each cytosine

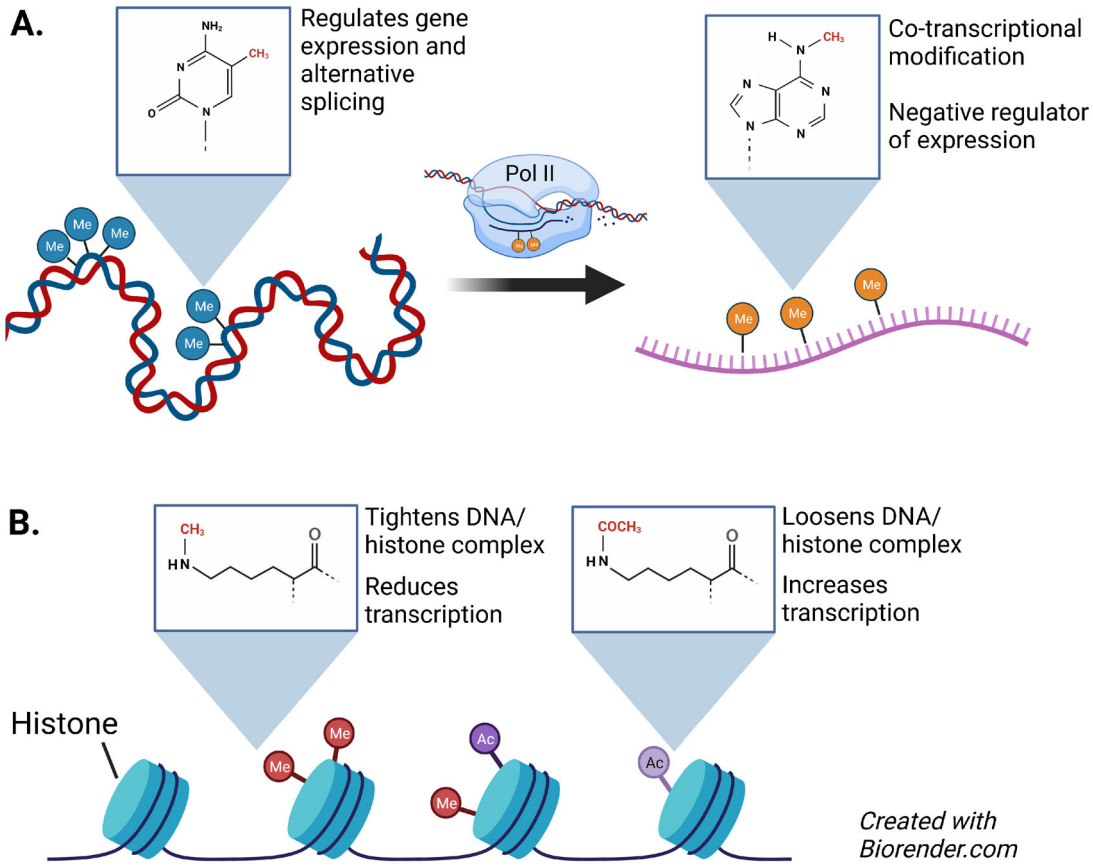

Figure 14. Major epigenetic modifications. (A) DNA methylation marks (methylated cytosine, or 5mC) and RNA methylation marks (N6-methyladenosine, or m6A). (B) Histone methylation and acetylation.

in the genome. Bisulfite sequencing is widely employed in epigenetics research to elucidate the role of DNA methylation in gene regulation and various diseases.

There are several strategies to study DNA methylation marks, including (1) whole genome bisulfite sequencing (WGBS), (2) reduced representation bisulfite sequencing (RRBS) using restriction digestion of the genomic sequences, or (3) targeted DNA methylation analysis using PCR. In honey bee research, WGBS is most often used, and is covered below.

5.2.1.1. Considerations

- Input: WGBS requires high-quality and high-quantity DNA samples. It is important to have enough DNA to generate sufficient sequencing depth to detect methylation at low-frequency sites. Caution must be taken because the sample processing can degrade DNA to some degree. Sufficient depth can be achieved by using large numbers of sequencing reads, or by using high-throughput sequencing platforms.
- Sequencing: Bisulfite-treated DNA needs to be converted into a sequencing library before it can be sequenced. This typically involves fragmenting the DNA, end-repairing the fragments, and then ligating adapters for sequencing.
- Integration: WGBS data is often compared with other epigenetic data such as ChIP-seq and RNA-seq to understand the functional relevance of DNA methylation.
- Validation: It is important to validate the WGBS data using other techniques such as pyrosequencing, bisulfite pyrosequencing, or methylation-specific PCR (MSP).

5.2.1.2. Materials

- Large equipment: Thermal cycler for PCRs, downstream sequencing platform such as Illumina MethylationEPIC Array, Single Cell Methyl-Seq, Agilent SureSelect MethylSeq (Agilent Technologies Inc.), or IDT xGen™ Methylation-Sequencing technologies
- Kits for bisulfite conversion and genomic library prep such as the EZ DNA Methylation Kits, generally including the M-Binding Buffer, M-Wash Buffer, L-Desulphonation Buffer, M-Elution Buffer, and Bisulfite conversion reagent, spin columns, and collection tubes.

5.2.1.3. Methods

5.2.1.3.1. Wet lab processing

1. Extract total genomic DNA (1 μg per group) from bulk tissue of interest using a total DNA extraction kit (for example, DNeasy Blood & Tissue Kit (QIAGEN), Maxwell® RSC Tissue DNA Kit (Promega), Quick-DNA Tissue/Insect Kit (Zymo Research)).
2. Use the EZ DNA methylation kit (Zymo Research), or equivalent, to perform the bisulfite treatment on all the genomic DNA samples following the manufacturer's instructions.
3. Bisulfite conversion reagent (130 μl) is added to a DNA sample (20 μl, DNA quantity range is between 100 and 2 μg) in a PCR tube.
4. Place the mix on the thermal cycler and run the following protocol: 98 °C (8 min), 54 °C (60 min), then 4 °C for overnight (no more than 20 h).
5. Add the converted DNA to a spin column which has M-Binding Buffer pre-added (600 μl).
6. Mix the sample and the buffer well, centrifuge for 30 s at >10,000 g.
7. Add M-Wash Buffer (100 μl) and repeat centrifugation
8. Add L-Desulphonation Buffer (200 μl) to the column and incubate 15-20 min (room temperature) and repeat centrifugation.
9. Add M-Wash Buffer (200 μl) to the column and repeat centrifugation. Repeat the wash again.
10. Transfer the column to a microcentrifuge tube (1.5 ml) and add 10 μl M-Elution Buffer, water, or TE buffer (pH value is more than or equal to 6.0) to the column center. Centrifuge 30 s at full speed to collect the DNA.

5.2.1.3.2. Sequencing and quality check.

1. Sequence the DNA on your preferred sequencing platform following the manufacturer's guidelines.
2. Check the sequencing data for quality control (QC), processed to align read to the most updated *Apis mellifera* genome (see Section 5.2.4). Several key QC measures to consider include the sequence quality, mapping efficiency as general sequencing process, coverage, and the bisulfite conversion efficiency. The coverage of each base should be at least 20×, and ideally higher.

TIP: If there is a high duplicate rate, it can indicate issues with library preparation or sequencing. It is generally recommended to have a duplication rate of less than 10%. It is also important to check the efficiency of bisulfite conversion, which can be done by comparing the proportion of C to T transitions in the WGBS data with the proportion of C to T transitions in the input DNA.

5.2.2. Methylated DNA immunoprecipitation-sequencing (MeDIP-seq)

Methylated DNA immunoprecipitation sequencing (MeDIP-seq) is a tool to identify regions of the

genome that are methylated. It is based on the principle that methylated DNA can be selectively pulled down from a sample using an antibody that recognizes 5-methylcytosine 5mC. Basically, the DNA is fragmented and denatured, then incubated with an antibody that specifically binds to 5mC. The methylated DNA is then immunoprecipitated using magnetic beads or protein A/G. The immunoprecipitated DNA is then sequenced, and the resulting data are analyzed to identify regions of the genome that are methylated.

5.2.2.1. Considerations. MeDIP-seq is a powerful technique that can be used to identify methylated regions in a genome-wide manner, and it also allows the detection of low levels of methylation. However, it has some limitations, such as the need for a high-quality antibody that recognizes 5mC and the potential for cross-reactivity with other modified forms of cytosine.

MeDIP-seq data can also be compared with other epigenetic data such as ChIP-seq and RNA-seq to understand the functional relevance of DNA methylation. The MeDIP-seq procedure described here is referenced from (Li et al., 2011) and (Shi et al., 2013).

5.2.2.2. Materials

- DNA template
- Genomic DNA extraction kit
- A Quick Ligation kit (QIAGEN)
- MinEluteH PCR Purification kit (QIAGEN)
- IP buffer (Tris-HCl 10 mM with pH 7.5, NaCl 280 mM, EDTA 1 mM)
- T4 polynucleotide kinase
- T4 DNA polymerase
- Anti5-methylcytosine mouse monoclonal antibody (Calbiochem)
- Dynabeads Protein G and Protein A

5.2.2.3. Methods

1. DNA is isolated by phenol-chloroform extraction, and fragmented (around 200–350 bp), and extracted by a gel extraction kit. The recovered DNA is first 5′ and 3′ end blunting, phosphorylating and repairing by T4 Polynucleotide Kinase and T4 DNA Polymerase.
2. After adding the ATP in the 3′ end, an Illumina sequencing primer adapter is ligated to the DNA using the Quick Ligation kit.
3. DNA is recovered by MinEluteH PCR Purification Kit (QIAGEN) and used for immunoprecipitation (IP).
4. IP DNA (4 mg) is then heat denatured (5 min), then is incubated with 32 mg anti5-methylcytosine mouse monoclonal antibody in a 400 ml of IP buffer for 30 min at 4 °C.
5. The Dynabeads Protein G (100 ml) and Protein A (Dynal) are added to the mix for incubation of 5.5 hours at 4 °C.
6. After immunoprecipitation, the DNA is amplified by PCR with specific Illumina sequencing primers. Illumina platforms such as HiSeq (discontinued), NextSeq, or NovaSeq can be used for sequencing, depending on the cost and coverage needs.

5.2.3. Data processing and analysis

The analysis of BS-seq data involves several steps, including data preprocessing, alignment, methylation calling, and downstream analysis. These steps can vary depending on the specific research question and the software tools used.

5.2.3.1. Software recommendations. The first step is to align the preprocessed reads to the reference genome. The most commonly used aligner for BS-seq data is Bismark, which is able to align bisulfite-converted reads to the reference genome (Krueger & Andrews, 2011).

After alignment, the next step is to call methylation levels at each cytosine position in the genome. There are several different methylation calling tools available, such as MethPipe (Song et al., 2013), DSS (Feng & Wu, 2019), and MethylSight (Biggar et al., 2020). The model-based analysis of chIPseq (MACS) approach can also be used to measure the methylation levels in the *A. mellifera* genome (Neary & Carless, 2020; Shi et al., 2013).

Once the methylation levels have been called, the next step is to perform downstream analysis to identify differentially methylated regions (DMRs), annotate them with functional elements and explore relationships with other data types such as transcriptomics. There are various tools available for this purpose, such as MethylKit (Akalin et al., 2012), DMRcate (Peters et al., 2021), and MethylSeekR (Burger et al., 2013). These tools use different algorithms to call methylation levels and correct for sequencing errors. The results can be visualized using various tools such as IGV, methylKit and Bioconductor packages such as BSgenome and BSseq.

5.2.3.2. Data repository. Like all sequencing data, DNA methylation or RNA methylation data should be deposited in a public database. For example, NCBI Sequence Read Archive (SRA) data portal is a public database for scientists to deposit their raw sequencing data (Section 11).

5.2.3.3. Statistical analysis. Statistical analysis of BS-seq data typically involves the estimation of methylation levels at individual cytosine positions or at

predefined regions of interest, such as gene bodies in the case of honey bee genome. The estimation of methylation levels is usually based on the proportion of methylated reads, which are reads that have a cytosine-to-thymine (C-to-T) change due to bisulfite conversion (Foret et al., 2012).

Once methylation levels have been estimated, several statistical methods can be used to identify differentially methylated regions (DMRs) between different samples or groups. Commonly used statistical thresholds for differential methylation are a percent methylation difference larger than 10% and $q < 0.01$ (Akalin et al., 2012). A Bayesian framework can also be used to estimate the probability of different methylation states at each cytosine position and to identify DMRs.

After DMRs have been identified, several downstream analyses such as functional annotation of DMRs, correlation of DMRs with other epigenetic or transcriptomic data, pathway analysis, or clustering of DMRs to identify patterns of co-regulation can be performed.

5.3. Epitranscriptomics: RNA methylation of m6A

Besides the chemical modifications on the DNA, RNA molecules can also undergo more than 100 different dynamic chemical modifications that impact gene expression (Roundtree et al., 2017; Sánchez-Vásquez et al., 2018). The study of RNA methylation so far has mainly focused on the N6-methyladenosine (m6A) modification, which occurs in mRNAs of eukaryotes and represents about 80% of all RNA methylation in the honey bees (Wang, Xiao, et al., 2021). In honey bees, recent discoveries have linked m6A methylation to larval development and caste differentiation, highlighting its significance in *Apis* biology (Bataglia et al., 2021; Wang, Xiao, et al., 2021). Epitranscriptomics is the field that focuses on studying all types of mRNA mRNA modifications, including m6A. The protocol below describes how to quantify total m6A methylation across all transcripts, but more advanced methods must be used to actually identify the sites of methylation.

5.3.1. Considerations for testing global RNA methylation of m6A

- This test requires a precise quantity of RNA input; therefore, it is crucial to have the equipment to quantify the RNA concentration in your RNA extraction sample. Nanodrop spectrophotometers or Qubit fluorometers can both provide accurate concentration information about your sample, but the latter is recommended for accuracy.
- The EpiQuik m6A RNA Methylation Quantification Kit (EpiGentek) can be used to quantify total m6A methylation globally across a complete RNA sample. The range of the RNA quantity is 100–300 ng. See Evans et al. (2013) for sample handling and RNA extraction protocols.

5.3.2. Materials

- Large equipment: Microplate reader capable of reading absorbance at 450 nm, 37 °C incubator
- Kits: EpiQuik m6A RNA Methylation Quantification Kit (EpiGentek) including negative and positive RNA controls, capture antibody, detection antibody, wash buffer, binding solution, enhancer solution, developer solution, stop solution, standard control
- 8-Well assay strips
- Distilled water
- 1× TE buffer with pH 7.5–8.0
- Adjustable pipette
- Aerosol-resistant filtered pipette tips
- 96 Well plates
- 1.5 ml Microcentrifuge tubes
- Plate sealing film or parafilm M

5.3.3. Procedure

1. Follow the kit manufacturer's instructions for RNA binding, antibody binding, and RNA quantification.
2. After color development, measure absorbance at 450 nm.
   a. For relative quantification, the percentage of m6A in total RNA can be calculated based on the average OD450 of negative control, average OD450 of positive control, average OD450 of sample, the amount of input sample RNA in ng, and the amount of input positive control in ng.
   b. For absolute quantification, a standard curve can be generated based on the m6A standard controls (0.01, 0.02, 0.05, 0.1, 0.2, and 0.5 ng/ul). With the standard curve, the amount of m6A in the total RNA sample can be calculated using sample OD minus negative control OD, then divided by the slope. The proportion of m6A in the samples can be calculated using m6A amount in ng divided by input quantity (for example 200 ng).

5.3.4. Identifying methylation sites

Although global m6A methylation is valuable for assessing large-scale methylation states, it is beneficial to understand which specific residues are modified on a transcript. Most RNA extraction techniques retain the methylation state on the mRNA transcripts, and there are several ways that the m6A sites

may be identified. m6A-seq uses ribose-sensitive nuclease digestion and high-throughput sequencing (Dominissini et al., 2013). MeRIP-seq (Meng et al., 2014) is another approach which uses m6A-specific antibodies to pull down m6A-containing RNAs and then sequences them.

A final option is Nanopore MinION sequencing. Using this technique, you can glean which specific RNA transcripts have m6A methylation as well as which residues are modified. Nanopore MinION direct RNA sequencing (dRNA-seq) provides long reads (10,000–30,000 base fragments being common) and, in some cases, can provide reads of entire transcripts. The advantage of dRNA-seq is that one sequencing procedure can include information of both transcriptome and m6A methylome modifications of RNA. See Section 6.2 for more information on direct RNA sequencing.

5.3.5. Software recommendations

For m6A RNA methylation from the Nanopore dRNA-seq data, one new tool for identifying differential RNA modifications is called xPore (Pratanwanich et al., 2021). xPore uses a machine learning-based approach to predict the presence and location of RNA modifications, including m6A, within the dRNA-seq reads. It also allows for the quantification of modification levels, and it can also be used to identify differentially modified sites between different samples or conditions. xPore's ability to handle long read lengths of direct-RNA sequencing data allows for accurate detection of m6A modifications in introns and exons, as well as in regions spanning multiple exons, resulting in higher resolution and more comprehensive coverage of m6A sites. It allows for high-resolution and comprehensive detection of m6A modifications, which can help to improve our understanding of the role of m6A in regulating gene expression.

One of the main challenges in the analysis of m6A site data is distinguishing true m6A modifications from sequencing noise. To address this, many tools use a threshold-based approach to call m6A sites, which involves setting a threshold for the minimum number of reads or the minimum proportion of methylated reads required to call a site as methylated. The choice of threshold will depend on the sequencing depth and the desired level of specificity and sensitivity. Similar to other transcriptomic analyses, multiple hypothesis testing must be accounted for (by using a Benjamini–Hochberg correction, for example).

5.4. Chromatin organization and histone modifications

Among a variety topics of epigenomics, transcriptional activities occurring at tissue and cellular levels are regulated by transcription factors (TFs) (Li-Byarlay et al., 2013; Qiu et al., 2013; Spitz & Furlong, 2012). TFs work with cis-acting regulatory elements (CRE) for cell regulation OD promoters, epigenetic patterns, and genome structure (Stadhouders et al., 2019).

A histone is an octomeric (8-subunit) protein complex which is a crucial component of chromatin. These complexes serve as a spool that genomic DNA is wrapped around. Post-translational modifications determine how tightly the DNA is associated with histones. Methylation of lysine 9 on the third subunit of the histone proteins (denoted H3K9me3) results in tighter association of the DNA to the histones, and limits the ability of transcription factors to access the DNA for transcription. Acetylation of several amino acids causes the DNA and histones to be associated more loosely, making the DNA more accessible to transcription factors and therefore more likely to be transcribed. Modification of the histone proteins provides an additional mechanism of gene regulation, beyond the binding of transcription factors to the promoter region of a gene.

5.4.1. Chromatin immunoprecipitation sequencing and transcription factor binding motifs

Chromatin immunoprecipitation sequencing (ChIP-seq) is one way that sites of histone modifications can be mapped to the genome. Like many modern molecular techniques, existing protocols and guidelines can be, more or less, directly adapted for use in honey bees with little modification (e.g., see Nakato and Sakata (2021)).

5.4.2. Hi-C & chromatin conformation

The Hi-C technology that captures genome-wide chromatin interactions and structure method has been already discussed in Section 3.3.1.1. Beyond genome assembly improvement, HI-C has identified that the variation of the genomic structure of *A. mellifera* is associated with the variation of phenotype (Jin et al., 2023), and is highly regulated with gene activities or adaptations in different environment (Kirkpatrick & Barton, 2006; Wallberg et al., 2017). In Drosophila fruit flies, or Heliconius butterflies, chromosomal inversions is related to adaptation to environmental adaptation (Joron et al., 2011; Krimbas & Powell, 1992). Future research using Hi-C technology is needed to obtain more information on the chromatin structure and how they interact with gene regulations.

5.4.3. Chromatin accessibility and transcriptional factor motifs

The chromatin accessibility is the percentage of time any given fragment of the genome is occupied by a

nucleosome. Although not necessarily heritable, it plays an important role in gene regulation and can change in response to environmental cues (Turner, 2008). Three assays used to study chromatin accessibility are: (1) assay for transposase accessible chromatin by sequencing (ATAC-seq), (2) DNase I hypersensitive site -seq (DNase-seq), and (3) micrococcal nuclear sequencing (MNase-seq). High read depth in these assays can reveal regions that where TFs bind (Tsompana & Buck, 2014).

Recent research using ATAC-seq, RNA-seq and ChIP-seq has identified many regulatory regions, including accessible chromatin regions, nucleosome occupancy, and specific patterns of TF gene networks in the genome of queen, worker, and drone adult bees (Lowe et al., 2022; Zhang, Li, et al., 2023). Another ATAC-seq and RNA-seq study comparing the TFs among the brains of foragers, newborn workers, and nurses revealed different regulatory landscape and new interactions within the transcriptional regulatory network underlying the division of labor (Fang et al., 2022). Previous meta-analyses and cis-metanalysis tools also revealed that transcriptional regulatory mechanisms that are underlying the behavioral maturation of honey bee workers (Ament, Blatti, et al., 2012). The genome-wide scanning and gene regulatory network activity also identified TF motifs in the honey bee brains at the colony and individual levels, and how they are associated with the regulation of social behavior and its evolution (Jones, Rao, et al., 2020; Rittschof et al., 2014; Shpigler et al., 2019).

Additional ChIP-seq experiments have informed on transcriptional factor binding sites of Vitellogenin gene, which affect the immunity and behavior of honey bees (Salmela et al., 2022). Other transcription factors such as usp, ubx, Mblk-1/E93 are also critical for phenotypic plasticity and development of honey bees and other insects (Ament, Wang, et al., 2012; Matsumura et al., 2022; Prasad et al., 2016; Yan et al., 2014). While these cutting-edge methods are still rare in *Apis*, the few studies that exist provide very detailed protocols for performing ChIP-seq experiments of histone methylation in honey bee larvae, including the extensive optimization such as cross-linking time and buffer compositions (Wojciechowski et al., 2018).

5.4.4. Detecting histone modifications by mass spectrometry

Histone modifications, a key epigenetic mechanism, significantly contribute significantly to phenotypic dimorphism and caste difference in *A. mellifera* honey bees (Dickman et al., 2013; Jin et al., 2023; Zhang, Li, et al., 2023). Several studies have reported that histone modifications are associated with social behavior, evolution, development, and ecology of *A. mellifera* (Alghamdi & Alattal, 2024; Dickman et al., 2013; Wojciechowski et al., 2018; Zhang, Li, et al., 2023). Histone deacetylase inhibitors also play a role in caste determination in honey bees (Spannhoff et al., 2011).

Mass spectrometry-based proteomics can be used to detect and quantify modified histones. Since typical proteomics workflow will dissociate histones from the DNA with which they interact, this approach is not suitable for identifying DNA binding sites. However, the power of mass spectrometry is its ability to detect numerous types of covalent modifications, with dozens of potential post-translational modifications of histones (Huang et al., 2015). While a generic shotgun proteomics dataset will contain ions corresponding to modified peptides, signal intensity and data quality are improved if immunoprecipitation is conducted to enrich the peptide or protein sample for sequences carrying the desired modification. The resulting mass spectrometry data processing will differ as well, as the mass shift and affected amino acid residue associated with the modification will need to be specified. Sample preparation procedures are essentially as described in Section 8, but further details can be found in see Huang et al. (2015).

5.5. Applications and limitations

Honey bees have long been studied as a model system for epigenetics in insects. Tools for epigenetic analyses have enabled research of epigenetic mechanisms underlying the complex behaviors, development, gene regulation, diseases, ecology, and health of honey bees (Galbraith et al., 2015; Grozinger & Robinson, 2015; Grozinger & Zayed, 2020; Kucharski et al., 2008; Li-Byarlay et al., 2020, 2013). Specifically, 5mC DNA methylation is involved in caste determination and behavioral maturation, which are essential for regular colony functioning (Herb et al., 2012). Furthermore, m6A modifications have been linked to behavior, learning, and memory (Bataglia et al., 2021; Wang & Li-Byarlay, 2015). However, research on m6A modifications in honey bees is still in its early stages, and further studies are needed to elucidate the specific role of m6A in regulating gene expression and splicing in this model organism.

Many studies also show that these epigenetic changes are tissue-specific or change over time (Li-Byarlay et al., 2020); therefore, genes with differential methylation patterns could be used in the future as biomarkers for environmental stress. In human cancer studies, for example, DNA methylation can be used to assist with diagnosis (Shames et al., 2007).

However, the study of epigenetics and epigenomics can be complicated by tissue-specific effects and temporal dynamics during different developmental stages of organisms (Li-Byarlay et al., 2020). Future research using multi-omic approaches that combine epigenomics with transcriptomics, genomics, and other 'omic tools may provide more comparative data. Until then, great care should be taken to control the age or developmental stage of samples used in these studies.

6. Transcriptomics

6.1. Introduction

Transcriptomics is the study of expressed transcripts, differential splicing patterns of messenger RNA, or chemical modifications on RNAs. What mechanisms program or influence genes expression, regulation, interaction? These questions can be answered holistically by sequencing the transcriptome and aligning it against an annotated reference genome.

Before NGS became accessible and affordable, northern blot techniques were most often used to detect and quantify mRNA transcripts. Quantitative reverse transcription-polymerase chain reaction (qRT-PCR) has also been used to answer similar questions. However, each of these techniques is practically limited to a small number of targeted genes. Transcriptomic techniques can identify nearly every transcript in an organism or tissue, and the relative quantities of these transcripts can be compared, providing valuable insights into the gene regulatory networks involved in a phenotype.

As honey bees were among the first insect genomes sequenced, the study of gene expression and regulations in *Apis* had the tremendous advantage of having an annotated reference genome as an anchor to map transcripts. At the time of publication of Evans et al. (2013), microarrays were the most commonly used high-throughput method of evaluating transcript abundances. Since then, RNA-seq has become a mainstay, with multiple sequencing platforms available, and even single-cell transcriptomics is possible. Here, we address these new techniques and how they can be applied to honey bee samples.

6.2. Sequencing technologies

Many of the same sequencing platforms used to sequence genomes can be used for transcriptomes as well (see Section 3: Genomic DNA sequencing, and (Evans et al., 2013)). The major differences between RNA-seq and DNA sequencing are (1) laboratory pre-processing conditions that require ultra-clean and low-temperature handling to limit RNA degradation and (2) the need to convert extracted sensitive RNA into double-stranded cDNA. Once transformed into cDNA, the material is stable and can be processed similarly to gDNA. Here, we describe specific considerations when applying this technology to transcriptome sequencing.

6.2.1. Considerations

- Always use proper microbiological aseptic techniques when working with the RNA and RNA-seq library to avoid degradation and cross-contamination. Transcript sequencing output will depend on the initial quality (RIN) and RNA degradation levels (Gallego Romero et al., 2014).
- Wear and change frequently disposable gloves to prevent RNase contamination and work as quickly as possible until cDNA is prepared.
- Keep tubes closed and on ice or cold block whenever possible.
- Be aware that as sequencing technology progresses, the platform chemistry can shift quickly. This can cause discordance in the data output and yield of the same RNA library sequenced (DeKayne et al., 2021).
- Downstream analysis such as differential gene expression analysis are sensitive to the number of replicates chosen per condition. A too low sample size may ignore natural variation and can be prone to false discoveries (Li, Janssens, et al., 2022; Squair et al., 2021).

6.2.2. Illumina sequencing (short reads)

Similar to genome sequencing, Illumina is the technology most often used for transcriptomic studies. Short reads are ideal for samples with low RNA yield, and it is relatively quick and inexpensive compared to other sequencing technologies. However, because this technology sequences many short strands with a fragment size ranging from 150 to 250 bp (Quail et al., 2009), reads from genes with multiple splice variants cannot always be mapped to a specific spliceoform. There are numerous resources describing short-read RNA sequencing and data analysis best practices (Conesa et al., 2016), which, combined with sample handling and RNA extraction protocols previously described specifically for honey bees (Evans et al., 2013), are directly applicable to honey bees. Here, we will focus on long-read sequencing and single-cell transcriptomics, but include updates to short-read sequencing covered in the previous *BEEBOOK* (Evans et al., 2013).

6.2.3. Third generation sequencing (long reads)

Third-generation sequencing enables the user to sequence contiguous and long reads from molecules of $>$10 kb on average, which minimizes some of the

computational workload needed to analyze full length transcripts, improves the accuracy of the alignments against the reference genome, and helps resolve ambiguous splice variants.

The main platforms for generating long reads include Oxford Nanopore Technology (ONT) and Isoform sequencing (ISO-seq) using Single-molecule, Real-time (SMRT) sequencing by PacBio. RNA extraction and sample handling for analysis by these platforms are essentially as previously described in Evans et al. (2013), with scrupulous attention to maintaining RNA integrity. Once RNA is extracted, the kit each sequencing platform manufactures, which are appropriate for their devices, should be used (for example, the Direct RNA or cDNA Sequencing Kit from ONT, or SMRTBell™ kit from PacBio).

6.2.3.1. Considerations for choosing a long-read platform.

- ONT has developed sequencing platforms to sequence RNA directly, without conversion to a cDNA intermediate. This minimizes the possibility of errors that could arise during a reverse transcription reaction, but the samples themselves are less stable leading up to sequencing.
- SMRT PacBio sequencing technology is an especially powerful platform because of its ability to generate long reads by reading a single molecule multiple times, which improves sensitivity.

6.3. Single-cell transcriptomics

RNA-seq has traditionally been done with whole tissue samples also called "bulk sequencing." However, recent advances in sequencing technology to target smaller input material have made it possible to sequence the transcriptomes of individual cells within a tissue sample while preserving the identity of their cell of origin to create a cell atlas. Single-cell transcriptomics is a relatively new technique developed in 2013 for biomedical purposes and only started to be applied to non-model organisms (Chen, Sun, et al., 2021). This field is uniquely suited to studying gene expression in known heterogeneous tissues and even identifying the spatial structure of cells in tissue downstream.

The approach takes a cell suspension of the tissue of interest, separates cells individually, and assigns an unique barcode to each mRNA transcript in the cell. This allows researchers to achieve incredible resolution within organs even creating cell atlas in mammals, fly and ants (Chen, Sun, et al., 2021; Li, Wang, et al., 2022; Li, Li, et al., 2022). Recently, Zhang, Wang, et al. (2022) used the technique to compare the transcriptomes of different types of neurons within the honey bee brain (e.g., Kenyon cells, optic lobe cells, olfactory projection neurons, etc.), initiating the first cell atlas in bees. Below is a general protocol for single-cell transcriptomics sample preparation adapted from Traniello et al. (2020).

6.3.1. Considerations

- Single-cell sequencing requires fresh samples that can readily be dissociated into a cell suspension.
- If it is not possible to use a fresh sample, the cell suspension (protocol described below) can be made fresh, then frozen and used later for sequencing.
- Cells embedded in excessive extracellular matrix run the risk of being damaged through dissociation and rendered unusable for sequencing.
- Different cell types from different tissues (for example brain versus fat body) will need more time to optimize the procedure for cell dissociation by testing different buffers or reagents.

6.3.2. Materials

- Large equipment: 10× chromium controller
- Pipette and tips
- Kits: Chromium Chip G Single Cell Kit (10× Genomics), Chromium Next GEM Single Cell 3′ Kit (10× Genomics)
- RNAse-free 1X PBS
- 1.5 ml microfuge tubes (nuclease-free)
- Trypsin
- Ethylene-diamine-tetracetic acid (EDTA)
- Phosphate-buffered saline (PBS)
- Dissection plate
- Dissection forceps and scissors
- Dissociation media
- Table centrifuge
- 0.4% trypan blue
- C-Chip disposable hemocytometer and a Compound microscope or a cellometer

6.3.3. Sample preparation procedure for single-cell sequencing

1. Rapidly dissect desired tissue in cold RNase-free 1X PBS and place in a 1.5 ml microfuge tube.
2. Add PBS with 0.25% trypsin to sample.
3. Incubate on ice for 30 minutes.
4. Gently dissociate cells by pipetting with 200 µl pipette tips.
5. Separate individual brain cells using a 10× Genomics microfluidics chip following the manufacturer's instructions.
6. Prepare single cell transcriptomics library utilizing the Chromium NextGen Single Cell 3′ technology according to manufacturer instructions.
7. Sequence single-cell libraries utilizing either Illumina or Nanopore sequencing.

6.4. Data handling and analysis

6.4.1. RNA-seq and differentially expressed genes (DEGs)

To identify differentially expressed genes in RNA-seq data obtained from a minimum of two conditions/treatments, here we provide an example analysis pipeline (Figure 15) using the aligner software STAR (spliced transcripts alignment to a reference) (Dobin et al., 2015), the software package Subreads (Liao et al., 2013) and a reference genome. Many bioinformatic pipelines exist for such analysis, some of which use open-source workflows and containers such as Nextflow to create community curated frameworks that facilitate and standardize DEGs detection (Ewels et al., 2020; Wratten et al., 2021). For example, nf-core/rnaseq pipeline offers an alternative to our proposed method, allowing for alignment to the genome with STAR or pseudo-alignment to the transcriptome with Salmon (Patro et al., 2017) or Kallisto (Bray et al., 2016) and adaptable for both bulk-tissue and single-cell transcriptomes. Additionally, it is important to mention that pseudo-alignment workflow might be better suited to users with a limited computational power and time as it achieves similar accuracy in terms of transcripts quantification than reference-anchor mapping and reads counts (Bray et al., 2016).

The resulting processed data can then be analyzed for differential expression using the R package, DESeq2 (Love et al., 2014). We optionally recommend that users compare and cross-validate the differential expression results with those of other packages, such as edgeR package (Liu et al., 2021), to verify the overlap of genes found up- and down-regulated. If the overlap is significant, this increase our confidence in identifying true differentially expressed genes, as both DESeq2 and edgeR varies in their statistical models to estimate differential expression. These software benefits from frequent updates and comprehensive tutorials are made available by the developer (e.g., DESeq2, and edgeR manual).

Step 1. After data pre-processing and quality control, map reads to annotated *Apis mellifera* genome.

- STAR can be run in the command line to map RNA-seq reads to a supplied reference genome. The most current version of the honey bee genome is Amel_HAv3.1 and can be downloaded on NCBI RefSeq.
- For running STAR in command line, generate a directory for the genome by using the following command:

```
mkdir starAligned
```

Step 2. Create the genome index with the following command:

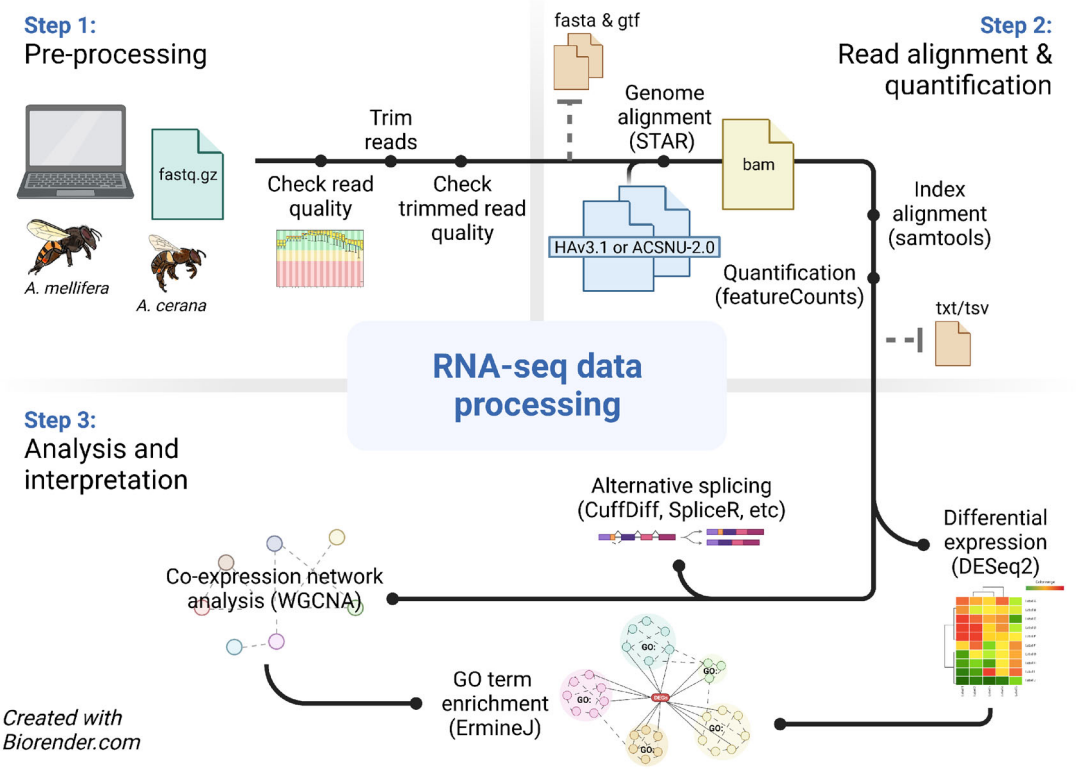

Figure 15. Example RNA-seq data analysis pipeline using a genome as anchor. Many software options are available for each step; only selected standard softwares are shown. Note key paces where data export may be desirable. Created with Biorender.

```
STAR –runThreadN 4
–runMode genomeGenerate
–genomeDir starAligned
–genomeFastaFiles/path/to/FASTAfiles
–sjdbGTFfile/path/to/GTF
```

The parameters were chosen for the following reasons:

`–runMode genomeGenerate` indicates we are in the mode to build genome index in the directory `–genomeDir starAligned.`

`–sjdbGTFfile/path/to/GTF` indicates that we want the annotation file in GTF format.

Step 3. Map the reads from the RNA-seq data onto the indexed genome, and place the counts into a table to be used later for the identification of differentially expressed genes. For paired reads, both the forward and reverse read will need to be provided as input. Because the FASTQ files are zipped, `readFilesCommand gunzip -c` must be used to unzip the files to be accessed by the mapping program. This script generates the aligned reads in a BAM file to be used by featureCounts.

TIP: Further questions related to the application of STAR can be found in the STAR manual.

```
STAR - -runThreadN 4
- -readFilesIn/path/to/genomeDir
- -readFilesIn/path/to/fastq1r1.gz,/path/to/fastq2r1.gz,\
/path/to/fastq1r2.gz,/path/to/fastq2r2.gz
- -genomeDir starAligned
#- -quantMode TranscriptomeSAM GeneCounts
- -outFileNamePrefix sampleName
- -outSAMtype BAM SortedByCoordinate
- -readFilesCommand gunzip -c
```

`–runThreadN 4` indicates that we run the mapping process using 4 threads.

`–genomeDir` indicates where the genome index is located.

`–outSAMtype BAM SortedByCoordinate` indicates that the output should be in BAM format and sorted by coordines.

Alternatively to featureCounts, we can also use directly the `–quantMode TranscriptomeSAM GeneCounts` to produce two outputs, one with the Read Count for each gene and one with the gene aligned to the transcriptome only.

Step 4. Counting reads can be accomplished using the featureCounts program, part of the Subreads package, in the command line. Once you have loaded Subreads, running the following command quantifies the reads associated with each gene. When the program is supplied with multiple BAM files, it will combine those reads into a single output which can be easily used in the DESeq2 R package for differential expression analysis.

```
featureCounts -a/path/to/annotationFile
-o sampleNameCounts.txt
-g gene_id
sampleName1Counts.BAM
sampleName2Counts.BAM
sampleName3Counts.BAM
```

Step 5. Identify differentially expressed genes using DESeq2 in R. This program takes the BAM files generated by featureCounts, and metadata about the individual samples, and calculates significant changes in gene expression based on

6.4.2. Gene network analysis

Gene ontology analysis can reveal groups of elevated or repressed genes classified by functional similarity. However, gene ontology does not represent groups of genes which have common patterns of regulation within groups. To uncover analysis pathways related to common regulatory mechanisms (genes regulated by a common transcription factor, for example), whole genome co-expression network analysis (WGCNA) can identify mechanisms of regulation (Langfelder & Horvath, 2008). This program, run in R, calculates modules of genes with correlated expression. Then, considering metadata, identifies modules whose expression correlates with those modules. Genes within individual modules can then be exported from the program, and can be used for further analysis (such as gene ontology).

An alternative to WGCNA is the ASTRIX (Analyzing Subsets of Transcriptional Regulators Influencing eXpression) method developed on honey bee brain TRN model (Chandrasekaran et al., 2011) which has been specifically designed for transcriptional regulatory network inference and has been widely applied in honey bee research. Recent studies have used the ASTRIX method in various tissues to identify the key regulatory genes and pathways that contribute to phenotypic plasticity and adaptation in *A. mellifera* (Chandrasekaran et al., 2011; Jones, Rao, et al., 2020; Shpigler et al., 2017, 2019; Traniello et al., 2023, 2020).

6.4.3. Single-cell transcriptomics

Single-cell sequencing data are unique in that each transcript has an unique RNA sequence tag added during the sequencing step. This modifier allows each read to be mapped to the cell in which it was transcribed. With this information, a complete transcriptome can be assembled for each individual cell in a sample. This creates an additional step in analysis, because each transcript must be assigned a cellular identity, in addition to being mapped to the

genome. Software packages that can be used to analyze these datasets are described below.

6.4.3.1. 10× Genomics specific software. 10× Genomics has a platform available specifically for the analysis of 10× Genomics datasets, using the programs Cell Ranger and Loupe Browser. The types of analyses, access to the platform, and a description of the coding involved are found on the 10× support webpage: https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/tutorials/gex-analysis-nature-publication.

Cell Ranger contains analysis pipelines for analyzing single-cell transcriptomic data, including converting raw base call (BCL) files to FASTQ files, aligning reads to a reference, and performing differential expression analysis. It can either be run in the 10× Cloud, or Cell Ranger can be downloaded and run from a desktop.

After expression profiles are generated for individual cells in cell ranger, they are grouped into clusters based on the co-expression of genes. Loupe Browser takes files created in Cell Ranger's aggr pipeline and converts them to a spatial representation of individual cells based on their expression profiles. Loupe Browser is available for download and can be run on a desktop.

6.4.3.2. Third party software. Seurat is an R package specifically designed to align, map, count, and analyze single-cell transcriptomic data (https://cran.r-project.org/web/packages/Seurat/index.html). Using weighted nearest-neighbor analysis, Seurat takes transcriptomic data and creates clusters of cells based on their gene expression profiles (Hao et al., 2021).

6.5. Applications and limitations

RNA-seq, particularly with short reads, is an incredibly versatile tool for gaining high quality transcriptomic data. However, long-read RNA sequencing data align more accurately to the genome and can help to identify the complete length of isoforms that are differentially expressed or differentially spliced. Long-read sequencing is thus filling an important gap in our knowledge of isoform expression patterns in honey bees.

Single-cell transcriptomics is a relatively new field, and while the potential resolution of this type of sequencing data are unparalleled, current practices do not yet reach the full potential. For example, to obtain a diverse representation of different cell types from targeted tissue, pooled samples are often still used. To harvest enough cells (50,000–100,000) for a good coverage of honey bee brain samples, brains from several honey bees are pooled together (Traniello et al., 2023, 2020). In the future, a key objective should be to decrease the quantity of the input cell numbers. Ideally, the ultimate goal would be to obtain deep sequencing data from individual bees instead of relying on pooled bee samples. This approach would significantly enhance statistical power and unlock the full potential of single-cell sequencing, providing exceptional resolution. A further limitation is that few cell types can currently be recognized based on the gene markers available. To overcome this limitation, it is crucial to conduct more extensive studies on individual gene markers that can accurately represent a broader range of honey bee cell types beyond neurons and glia.

Transcriptomics and RNA-seq analyses can provide information on the expression, isoforms abundance, alternative splicing, and chemical modifications of genes. Additional information about the activity of proteins or other cellular processes is essential for understanding gene function, and can be added by designing additional experiments using functional genomics (Section 7) proteomics (Section 8) and metabolomics (Section 9).

7. Functional genomics and xenobiotic treatment

7.1. Introduction

By analyzing the entire genome, researchers can explore intergenic relationships that extend beyond monogenic, individual polymorphisms. Recent developments in sequencing technology, combined with publicly available computational tools, have made it possible to glean a huge amount of information from these datasets. However, the challenge lies in our ability to fully interpret the data, particularly in cases where gene functions are either unknown or poorly characterized. In fact, approximately one third of *A. mellifera* genes fall into this category, posing limitations on our understanding of their roles and contributions.

This gap in functional knowledge can be filled by experimentally manipulating genes of interest to directly examine their contribution to a specific phenotype. Gene expression can be manipulated through changes to DNA (e.g., using CRISPR/Cas9 technology) or knockdowns of RNA (RNAi). These techniques can provide researchers with causal information about their phenotype of interest to go along with correlative data from genomic analysis. This is a necessary step for validating gene functions and gene interactions.

7.2. CRISPR

Clustered regularly interspaced palindromic repeats and Cas9 (CRISPR/Cas9) gene editing is a relatively recent molecular tool for modifying the genome at very precise locations (Jinek et al., 2012). Knock-outs (removal of a large portion of a gene from the genome), knock-ins (insertion of an endogenous gene, a modified gene with altered expression, or a reporter), and targeted gene editing are all possible with the use of CRISPR. CRISPR has already been used to study the role of major royal jelly proteins (Hu et al., 2019; Kohno et al., 2016), taste (Değirmenci et al., 2020), development (Kohno & Kubo, 2018), neurodevelopment (Chen, Traniello, et al., 2021), and mechanisms of sex-determination (Roth et al., 2019; Wang, Lin, et al., 2021) in honey bees. It is not, however, the only method of manipulating gene expression; other approaches are discussed in McAfee et al. (2022).

CRISPR technology relies on a well-designed short guide RNA (sgRNA) sequence to target a specific region of the genome. This sgRNA must include the following components:

- A tracr RNA sequence, which creates a hairpin to act as a scaffold for Cas9 binding
- A crRNA (CRISPR RNA) sequence (17–20 bp) which is specific to the target DNA

Constructs can be designed manually or utilizing a web-based tool. Several such tools exist, but many do not currently query against the *A. mellifera* genome. However, the tool Cas-OFFinder has this option. In addition, the region of interest within the genome must have a PAM (protospacer adjacent motif) sequence. This serves as a guide for the actual site of cleavage, which is 3–4 bp upstream of the PAM sequence. The following protocol for germline gene editing was inspired by Chen, Traniello, et al. (2021) and Hu et al. (2019). Methods for egg collection and microinjection were originally developed by Beye et al. (2002).

7.2.1. Considerations

- Both Cas9 protein and sgRNA can either be purchased from vendors or expressed and purified in-house. The following protocol is for in-house expression. If purchasing components from vendors, skip to Section 7.2.1.3.
- This protocol describes injecting the gene editing components into honey bee eggs, but injecting into non-embryonic tissues (e.g., adult brains) is also possible.
- Egg injections typically have a high failure rate and injecting thousands of eggs will likely be necessary to obtain an adequate number of mutants. To collect large numbers of fresh eggs, we recommend using the Jenter egg collection kit, which allows eggs to be conveniently collected on removable plastic plugs, although other methods are possible (e.g., (Lee & Lee, 2019)). Condition queens by introducing them to the cages several days before injection day. You may need 3–5 laying queens to collect a sufficient number of fresh eggs.
- Check with your institution's Biosafety department for regulations around handling, maintaining, and disposing of genetically modified insects.

7.2.2. Materials

- Large equipment: Incubators for bacterial culture and injected eggs, microinjector (e.g., PLI100 Pico injector (Warner Instruments) or Femtojet 4i (Eppendorf)), dissection scope, micromanipulator, basic laboratory equipment (microfuge, vortex, water bath, etc.), and a Nanodrop (Thermo Fisher) or Qubit (Thermo Fisher)
- Kits: QIAprep plasmid purification kit (QIAGEN, or equivalent), T7 RiboMAX Express Large Scale RNA Production System (Promega), Monarch RNA Cleanup Kit (New England Biolabs, Inc)
- Basic E. coli transformation and culturing reagents and materials
- IPTG (IPTG (isopropyl β-d-1-thiogalactopyranoside)
- Ni-NTA Superflow resin column (QIAGEN)
- PD-10 column (GE Life Sciences)
- BsaI restriction enzyme and digestion buffer
- Cas9 storage buffer (20 mM Tris, pH 8, 200 mM KCl, 10 mM $MgCl_2$, 10% glycerol)
- Injection buffer (20 mM HEPES, pH 7.5, 300 mM KCl, 1 mM $MgCl_2$)

7.2.3. Methods for CRISPR/Cas9 gene editing of embryos

7.2.3.1. Generating Cas9 protein.

1. Obtain the plasmid pET-28b-Cas9-His (Addgene, Watertown, MA, USA) for Cas9 expression under the control of the lac operator. Follow standard protocols for transformation, culturing, and IPTG induction in *E. coli* cells.
2. Purify the His-tagged Cas9 from protein lysate with a Ni-NTA Superflow resin column.
3. Desalt the Cas9 using a PD-10 column.
4. Once eluted, Cas9 protein can be stored as a 50 μM solution in storage buffer at –80 °C.
5. For new batches, check Cas9 purity by gel electrophoresis.

7.2.3.2. Generating sgRNA.

1. Ensure that equipment, surfaces, and reagents are nuclease-free.

2. sgRNA cDNA can be inserted into a MiniGene plasmid (or equivalent plasmid for T7 *in vitro* transcription). Follow basic *E. coli* transformation and culturing protocols to increase plasmid copy numbers.
3. Extract plasmid and purify using a plasmid QIAprep kit (QIAGEN) or equivalent.
4. Linearize plasmid by digesting with BsaI (if using a MiniGene plasmid), or equivalent.
5. Conduct *in vitro* transcription on linearized plasmid using the T7 RiboMAX Express Large Scale RNA Production System.
6. Purify RNA with the Monarch RNA Cleanup Kit according to the manufacturer's instructions for purification and storage.
7. For new batches, check RNA integrity by gel electrophoresis and quantify using a Nanodrop or Qubit

7.2.3.3. Ribonucleoprotein assembly.

1. Combine a 1:2 molar ratio of Cas9 to sgRNA, at a final concentration of 5 µM RNP, in injection buffer.
2. This working solution can be split into 6 µl aliquots and frozen at −80 °C.
3. After thawing, dilute solution to 2.5 µM with injection buffer.

7.2.3.4. Egg collection and microinjection.

1. Train queen in queen cages (Jenter™) (Figure 16(A)) by intermittently caging her for one day at a time. This will also give the workers a chance to build out the comb.
2. On injection day, replace old egg plugs with fresh plugs. 2–4 h later, collect the freshly laid eggs by removing the plugs and fastening them to plasticine disks (Figure 16(B)).
3. Keep eggs warm (30–33 °C) at all times or development will be delayed. Store in an incubator unless actively injecting. Perform injections in a walk-in incubator if possible.
4. Load ~1 µl of RNP working solution into the injection pipet.
5. Inject approximately 300–400 pL into the egg. Research has shown that injecting into the anterior third of the egg is most effective (Otte et al., 2018).
6. Eggs should hatch about 72 h after injection. A few hours before hatching, prime the egg cups with a small amount of warmed royal jelly (Figure 16(C)). Remove cups with shriveled or deformed eggs.
7. Once hatched, follow methods described in "Standard Methods for Artificial Rearing of *Apis mellifera* Larvae" (Crailsheim et al., 2013).

7.3. RNA interference

RNA interference (RNAi) is a molecular tool used for transient knock-down (reduction) of gene expression (Brutscher & Flenniken, 2015; Wilson & Doudna, 2013). This technique takes advantage of endogenous molecular machinery that normally helps protect the host against RNA viruses, which replicate their genome via a double-stranded (ds)RNA intermediate. The presence of dsRNA within host cells thus signals viral infection and triggers a cascade of events leading to viral inactivation by the RNA-induced silencing complex (RISC).

Although RNAi evolved as an immune defense against viruses, the dsRNA that leads to the antiviral response does not necessarily need to be a viral sequence. RNAi against nearly any gene is theoretically possible and can be achieved simply through the introduction of short interfering RNA (siRNA; 20–25 bp) or long dsRNA (several hundred bp) to the organism – a technique that has proven to be

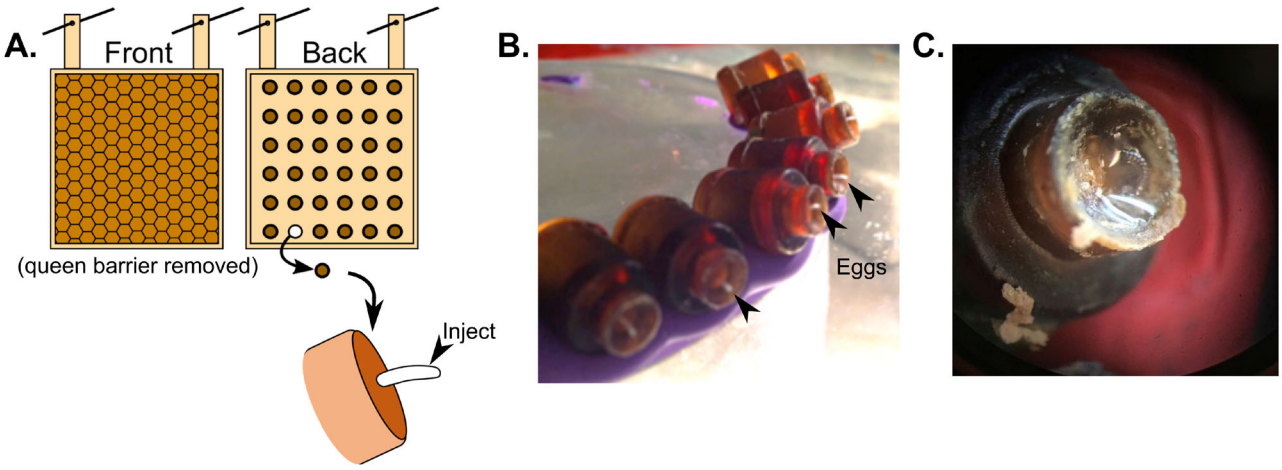

Figure 16. Examples of the Jenter egg collection system. (A) Cartoon diagram of an egg collection cassette (not to scale, real cassette has 110 plugs). Image adapted from McAfee et al. (2018) (CC BY 4.0). (B) Plugs with eggs fastened on modeling clay. (C) Newly hatched larva primed with royal jelly.

especially feasible in insects (Huvenne & Smagghe, 2010). Here we present current methods for conducting RNAi experiments in honey bees.

7.3.1. RNAi considerations

- A common method of introducing RNAi constructs into bees is via injection; however, this is highly invasive and creates a wound on the bee. This might be useful if modeling parasitization by mites, but in most cases, the wound creates undesirable trauma. The methods described here are less invasive.
- Feeding dsRNA to larvae or adults is also a non-invasive method. It has been covered already in Evans et al. (2013).
- Non-specific effects associated with RNAi are well-documented (Flenniken & Andino, 2013; Jarosch & Moritz, 2012; Nunes et al., 2013) and probably unavoidable, but the likelihood of observing severe off-target phenotypes can be reduced by cross-referencing the siRNA/dsRNA sequence to the honey bee genome. Only sequences unique to the target should be pursued. Moreover, negative control groups utilizing GFP dsRNA or scrambled dsRNA sequences are essential to distinguish gene-specific phenotypes from non-specific effects.
- Consider labeling the siRNA construct, either with a fluorescent tag (GFP or RFP) or digoxigenin (DIG) to visualize localization in tissue after treatment and check that the siRNA is correctly introduced to the target tissue.
- An alternative strategy for RNAi using symbiont-mediation is also documented (Lariviere et al., 2023)

7.3.2. Methods for nanoparticle-mediated RNAi

Honey bees breathe through spiracles, which are small holes in the abdomen connected to a tracheal network. In addition to feeding and injection, one way to introduce a foreign substance into a bee is through inhalation. Spiracles are approximately 200 μm wide; therefore, nebulized nanoparticles carrying RNAi constructs can enter the bee, assuming the nanoparticle-containing droplets or aerosolized particles are sufficiently small. Here we describe the general procedure for using perfluorocarbon (PFC) nanoparticles for delivering siRNA to bees, adapted from methods described in Li-Byarlay et al. (2013).

7.3.2.1. Materials

- Large equipment: humidity-controlled incubator
- Medical nebulizer (purchased from a medical health supply store)
- Collecting chamber with end cap (Bioquip*, Catalog #2820GA: 2820D) attached to the nebulizer *Bioquip is now closed permanently. Equivalent needs to be searched
- Nanoparticles that can carry small interfere RNAs (PFCs purchased from Thermo Fisher, for example)
- siRNA (follow methods outlined in Evans et al. (2013) for siRNA design, production, and purification, or purchase from a vendor)
- Acrylic or plexiglass cages for maintaining live bees in the laboratory and which are suitable for exposure to nebulized nanoparticles

7.3.2.2. Procedure.

1. To coat nanoparticles with siRNA, mix siRNA and add nanoparticles to create a solution with a ratio of 1 μM siRNA to 200 pM nanoparticles.
   TIP: Higher ratios may be used, but will not necessarily be more efficacious.
2. Select bees and maintain them in laboratory cages according to methods outlined in Williams et al. (2013).
3. Add the siRNA and nanoparticle solution to a nebulizer compressor machine. Live bees to be treated are housed in the collecting chamber. The solution is then sprayed on bees for ~5 min.
4. Allow bees to recover in a dark incubator held at 32 °C and at least 40% humidity for 96 h. Provide pollen paste (50% *w/v* honey, 50% *w/v* pollen) and sugar water (50% *w/v* sucrose in water) *ad libitum*.

7.4. Xenobiotic treatment

Treating bees with xenobiotic compounds is a quick and non-invasive means for modifying neurotransmitters and the activity of signaling pathways. In addition, it provides a convenient means for studying the impact of environmental chemicals (e.g., pesticides) on bee health. Most of the protocols below are described in Barron et al. (2007). Furthermore, additional types of compounds such as hormones and neurotransmitters can be used to treat bees and study their functions in behavior (Blenau & Baumann, 2016).

7.4.1. Xenobiotic treatment considerations

- For any application, it is important to consider the realistic and relevant dosage. One way to determine this is to conduct assays with different xenobiotic concentrations and observe any changes in health or behavior which indicate that the chemical is active in the bee.
- Depending on the goals of the experiment, the distribution of the chemical throughout the bee's body may need to be monitored to confirm that it enters the tissue of interest. This has been

done previously using radiolabeled chemicals (Barron et al., 2007) so that the radioactivity could be traced after administration.
- Always follow chemical safety guidelines when working with pesticides and other chemicals that could be hazardous to human health
- Many different drugs of interest can be taken up using this method, and requires very little reagent to see an effect. Serial dilutions will likely be necessary to generate solutions for application with precise concentrations.
- The appropriate controls such as injection with saline should also be considered.

7.4.2. Materials

- Large equipment: Dissection microscope, Hamilton syringe, flight cage, in-hive feeders or feeding stations
- Insect Ringer solution (0.125 M sodium chloride, 1.5 mM calcium chloride dihydrate, 5 mM potassium chloride, 0.8 mM sodium phosphate dibasic, pH 7.4, filter sterilized)
- High-purity xenobiotic compound
- Organic solvent or water for to produce xenobiotic stock solution, depending on solubility

7.4.3. Procedure

7.4.3.1. Thorax application. Different compounds may be soluble in different solvents. Acetone (Li-Byarlay et al., 2014) and dimethylformamaide (DMF) (Barron et al., 2007) have been previously described for topical applications. This is the simplest application method; however, actual absorption efficiency of most drugs through the cuticle is unknown.

1. Anesthetize bees on ice or otherwise immobilize them for treatment.
2. Apply 1 µl of solution containing the xenobiotic dissolved in organic solvent directly to the thorax of bees.
3. Continue to immobilize bees for at least 30 seconds after application to allow for absorption and solvent evaporation.
4. Analysis of phenotype (behavioral assay, dissection, genetic analysis) can be conducted 24 h after treatment.
5. The amount of time suggested here may be different depending on the treatment and phenotypes.

7.4.3.2. Injection. Injections allow very precise quantities of the drug to be directly administered to tissues or body sections of interest. This is especially useful for determining the action of a particular gene or signaling cascade in a specific location. However, this method requires specialized equipment, which the feeding methods do not. In addition, the person administering the drug must be incredibly careful when controlling the drug delivery to ensure that there is no damage to the bee in the course of the experiment.

1. Dilute the xenobiotic stock solution in insect ringer solution to the desired concentration.
2. Administer 1 µl of solution to each bee using a Hamilton syringe.
3. Inject the bee.
   a. For injections in the thorax, insert the needle at the base of the mesonotum to the right of the midline.
   b. For injections into the head, the solution should be applied to the median ocellus. To do so, anesthetize the bees and place them on a strip of duct tape, such that the head is immobilized. Remove the lens of the median ocellus with a micro scalpel and apply 1 µl of the solution to this area. Confirm that absorption occurs over several minutes.

7.4.3.3. Feeding individual bees. This method allows for control over the quantity of drug administered and allows for the researcher to control the time elapsed between the administration and any behavioral or genetic assay. However, administering drugs in this way can be very tedious and time consuming. It is also easier to damage bees in the course of collection, harnessing, and feeding.

1. Prior to feeding, starve bees for 1–4 hours.
2. Dissolve your drug of interest in a sucrose solution (50% *w/v*). TIP: Depending on the solubility of the chemical, a carrier solvent may be needed, such as lecithin or ethanol. If a carrier is used, it is crucial to take this into account when conducting vehicle controls.
3. Immobilize bees in a 1.5 ml microfuge tube that has a hole made in the conical end.
4. With a single bee in the tube with their heads towards the conical end opening, administer the drug in the sugar solution via pipette. Visually confirm that the bees are eating the solution. Bees can be fed up to 10 µl of solution in this way.

7.4.3.4. Flight cage feeding. Although feedings are typically done in the laboratory, it is possible to conduct pharmaceutical feedings on a larger scale. However, a major concern when feeding colonies in the field is the potential for contaminating the environment with the drug being administered. One way around this is to utilize a closed system, such as a

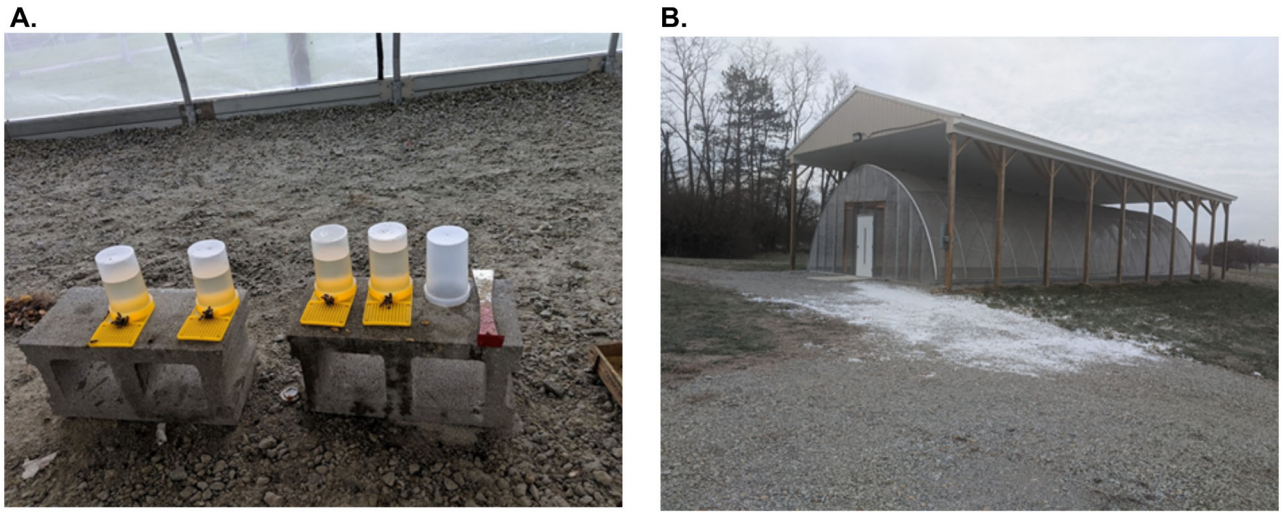

Figure 17. Setup for flight cage feeding. (A) Bees are provided with a sugar water solution with the compound of interest. (B) Flight cage enclosure is covered to protect from direct sunlight and rain.

flight cage, where no other pollinators will be exposed. In a flight cage system (Figure 17), the drug of interest can be administered in a sucrose solution (1:1 weight/volume) using in-hive feeders or feeding stations (Momowoa). Food coloring can also be mixed with the sucrose solution, so that its ingestion can be observed in the bees.

This method enables researchers to feed compounds to an entire honey bee colony, without exposing the drug to other animals or plants. This method requires the installation of a specialized facility (the flight house), which can be expensive and time consuming. Because the bees can only eat food that is provided to them, fresh sugar solution and pollen must be provided daily. Food that isn't replaced regularly might become moldy and toxic to the bees.

7.5. Applications and limitations

In honey bees, the genome is complex and many genes are not fully understood; indeed, over one third still are not functionally annotated (Walsh et al., 2022). Therefore, it can be difficult to know what particular genes control a trait or function, and RNAi is most commonly applied as a discovery tool in order to elucidate the roles genes play. There is also variability in responses of the knockdown effects. Factors such as genetic variation, environmental factors, and widespread off-target effects (Schulte et al., 2014), which are difficult to predict or control, sometimes making it challenging to replicate results.

Because of the haplodiploid system of sex-determination and complex mating strategy of honey bees, there is limited applicability to manipulate gene expression long-term (that is, to develop mutated lineages using CRISPR or transgenic techniques). Though there has been some success rearing transgenic queens (Schulte et al., 2014), those queens would need to mate with transgenic drones in order to establish a fully modified colony. Such an endeavor is exceedingly difficult and has not yet been achieved, but unmated queens can be stimulated to lay haploid eggs in microcolonies to yield genetically modified drones (Schulte et al., 2014). Moreover, ethical considerations of bee containment and risk of escape further inhibit widespread use of this technique.

The high efficiency of genetic modification by CRISPR, however, is making it increasingly feasible to avoid the need of creating a genetic lineage to produce large numbers of modified bees for experiments. Instead, embryos can be injected and reared *in vitro* in sufficient numbers to facilitate developmental studies (Roth et al., 2019). This method also poses virtually no risk of viable bees escaping into the environment. For these reasons, this seems to be the direction in which the field is headed, but studies on social phenotypes, such as the dance language or hygienic behavior, are not possible.

8. Proteomics

8.1. Introduction

While massively parallel sequencing enables high-throughput analyses of gene expression, transcripts are still one step away from the proteins that actually execute most biological functions. Gene and protein expression patterns are often not well correlated (Payne, 2015); therefore, transcriptomics and proteomics techniques are complementary. Diverse technologies have been used for proteomics over the years, including two-dimensional gel electrophoresis, mass fingerprinting, top-down proteomics, antibody microarrays, and shot-gun proteomics. Shot-gun

proteomics has become the dominant technique, representing the vast majority of proteomics work done currently, and will be the focus of the methods described here.

In this approach, proteins are first digested to peptides using proteases, then the peptides are ionized (acidified) and measured in a liquid chromatography-coupled tandem mass spectrometer (LC-MSMS) instrument. This type of instrument first measures the mass-to-charge ratio of a peptide ion, then fragments the ion and measures the mass-to-charge ratios of the fragments. These data can be used to identify the original peptide sequence, and bioinformatics tools are used to infer which proteins were present in the sample and in what quantities. Proteomics has historically lagged behind transcriptomics in terms of coverage and sensitivity (transcriptomics datasets typically quantify tens of thousands of genes, whereas proteomics datasets typically quantify several thousand proteins) (Timp & Timp, 2020), but improved liquid chromatography systems, instrumentation, and software have made proteomics competitively powerful (Aebersold & Mann, 2016).

Despite these exciting advances, it is still challenging to achieve rich proteomics datasets for honey bees relative to model species (McAfee et al., 2016). As with many non-model organisms, honey bees have undergone fewer iterations of genome annotation refinement than humans and model species, which hinders the ability of mass spectrometry data processing algorithms to assign spectra to peptide sequences. In a typical workflow (Figure 18), spectra can only be matched to known peptide sequences; therefore, proteome coverage is inherently sensitive to how precise and complete the genome annotation is. At the time of writing, the highest honey bee proteome coverage yet published in a single study is 4,604 unique protein groups (McAfee et al., 2021), obtained from unfractionated shot-gun analysis of $n = 28$ samples of eggs, but identifications more typically range from 1,000 to 3,000 protein groups (McAfee et al., 2016). Here, we outline the experimental procedure and data processing steps used to obtain this high-coverage dataset, although there are many potential variations on the protocol to fit different needs (e.g., alternate lysis buffers, digestion buffers, precipitation methods,

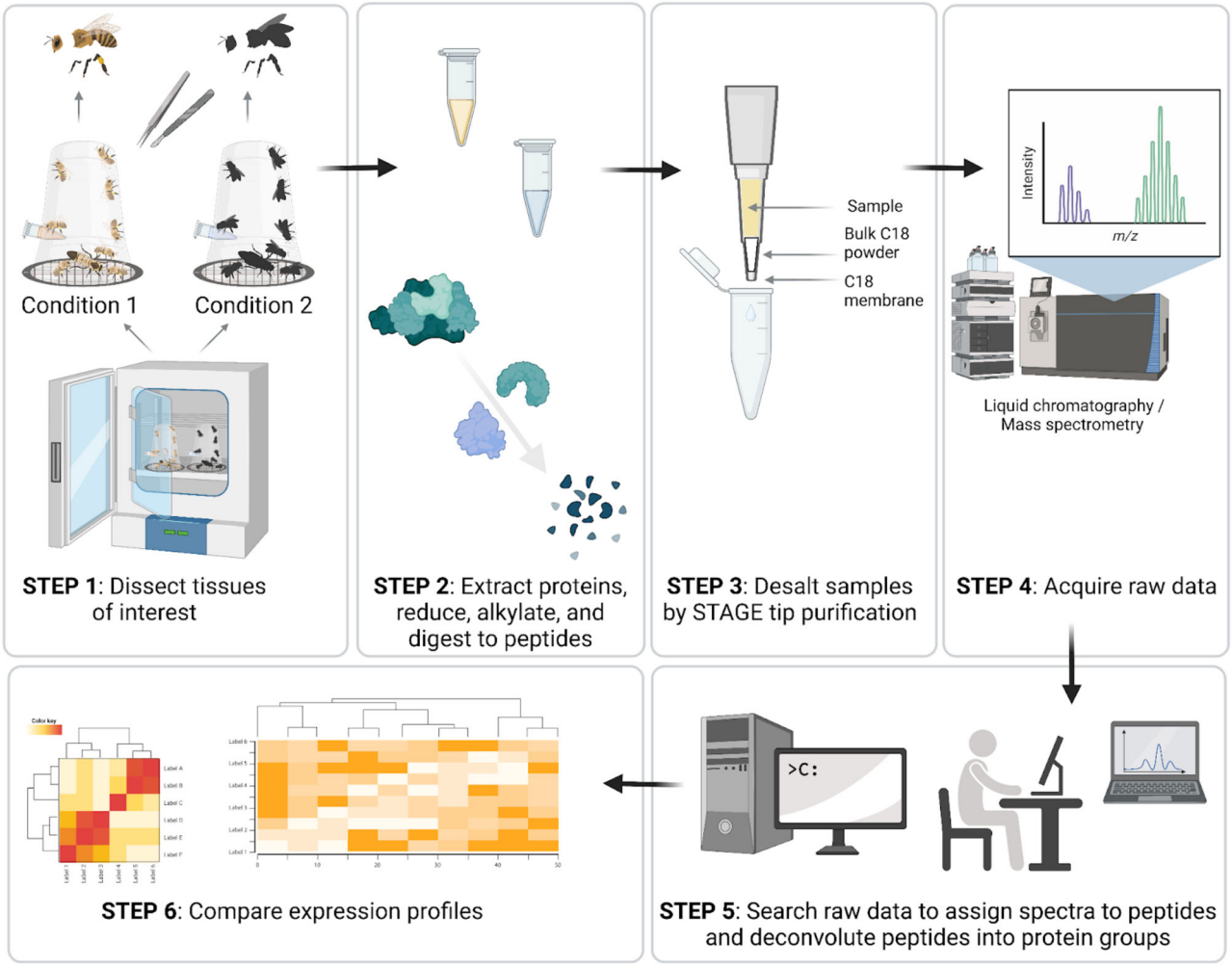

Figure 18. Schematic of a typical shot-gun proteomics workflow.

peptide desalting approaches, chromatography systems, etc.). The following methods work well for most honey bee proteomics samples.

8.2. Standard methods for shot-gun proteomics sample preparation

8.2.1. Considerations

8.2.1.1. General.

- Depending on the type of tissue, expect between 1 and 10% of the total wet mass to be made up of protein.
- Aim to digest 10–30 μg of protein per sample – more if performing additional sample enrichment or fractionation steps (e.g., high-pH reverse phase fractionation, phosphopeptide enrichment, size exclusion chromatography, etc) ahead of LC-MS/MS (reviewed elsewhere (Hedrick et al., 2015)).
- Mass spectrometer and liquid chromatography parameters are reviewed elsewhere (Savaryn et al., 2016).

8.2.1.2. Sample handling.

- Always use gloves and keep the workspace free from dust to prevent keratin contamination. Do not wear wool clothing.
- Adult honey bees and larvae can be coated with unwanted substances (e.g., honey, pollen, or royal jelly). Wash the samples by gently vortexing one or more times in phosphate-buffered saline (1x PBS) prior to lysis to help remove contaminants.
- If preparing fatty tissue samples, such as fat bodies, avoid retaining the top layer of fat after clarifying the lysate.
- If preparing gut samples, consider that there could be many plant, fungal, and bacterial proteins present inside. These could be washed out or included in the analysis, depending on the purpose of the experiment. But, if included, their protein sequences must be added to the search database in order to identify them.
- Guanidinium chloride will denature enzymes and must be removed from the sample or diluted prior to digestion. Protease inhibitors are not necessary during sample lysis if using a guanidinium chloride extraction buffer, but may be considered if a non-denaturing buffer is used.
- Mass spectrometry is not compatible with polymers (e.g., polyethylene glycol, PEG) and detergents (e.g., sodium dodecyl sulfate, SDS). If such reagents come in contact with the sample, they must be removed prior to mass spectrometry analysis. Consult Keller et al. (2008) for a list of common contaminants.
- Mass spectrometry data analysis is not compatible with disulfide bonds between peptides. Disulfide bonds must therefore be reduced and alkylated to block the reactive sulfur.

8.2.1.3. Reagent handling.

- All reagents should be mass spectrometry grade.
- When preparing reagents with undiluted strong acids, use glass pipettes for dispensing, as polymers and plasticizers can leach from plastics in contact with strong acids.

8.2.2. Materials

- Large equipment: sample homogenizer (e.g., Precellys 24), speed-vac, centrifuge, bath sonicator (optional), liquid chromatography system coupled to a mass spectrometer (see Section 8.3 for more details)
- Homogenization tubes (e.g., 2 mL DuraTubes) and ceramic homogenization beads
- C18 membrane or inert material for a column frit
- Bulk C18 powder (e.g., ReproSil-Pur 120 C18-AQ, 2.4 μm)
- Lysis buffer (6 M guanidinium chloride, 100 mM Tris, pH 8.0) – store at room temperature (RT) for several days, or 4 °C for weeks or months
- 100% Acetone (pre-chilled to −20 °C)
- 80% Acetone in water (pre-chilled to −20 °C)
- Bradford assay reagents (store at 4 °C)
- Step 1 digestion buffer (6 M urea, 2 M thiourea, 100 mM Tris, pH 8.0) – make fresh
- Step 2 digestion buffer (50 mM ammonium bicarbonate, pH 8.0) – make fresh
- Endoproteinase lys-C, mass spectrometry grade (see manufacturer's protocol for storage)
- Porcine-modified trypsin, mass spectrometry grade (see manufacturer's protocol for storage)
- Dithiothreitol (DTT; 0.5 μg/μl in water) – store at −20 °C
- Iodoacetamide (IAA; 1 μg/μl in water) – store at −20 °C, sensitive to light
- Buffer A (0.2% trifluoroacetic acid in water) – store at RT
- Buffer B (0.1% formic acid, 80% acetonitrile in water) – store at RT
- Elution buffer (0.1% formic acid, 40% acetonitrile in water) – store at RT
- Resuspension buffer (0.1% formic acid, 2% acetonitrile in water – store at RT

8.2.3. Proteomics methods

8.2.3.1. Lysis and precipitation.

1. Place the tissue sample (1 – 100 mg) in a homogenization tube with four ceramic beads and cold lysis buffer (maximum ratio: 100 mg tissue per mL). Work on ice.

2. Homogenize the sample (3 × 30 s, 1 min on ice in between, maximum frequency 6,000 rpm). TIP: The sample should be pulverized with no large pieces left in the solution. If you do not have a bead mill homogenizer and have a large sample quantity, the sample can be ground in liquid nitrogen with a mortar and pestle, then transferred to a tube containing lysis buffer.
3. Quick spin, then transfer lysate to a new tube. Spin sample at 16,000 g for 10 min (4 °C) to remove debris. Transfer supernatant to a new tube. Repeat if necessary.
4. Precipitate protein from the clarified lysate by adding 4x the sample volume of ice-cold 100% acetone (−20 °C, overnight - final solution should be 80% acetone).
5. Spin sample at 5,000 g (15 min, 4 °C) to pellet precipitated protein. Discard the supernatant and wash the pellet with 500 µl ice-cold 80% acetone. For large pellets, disrupt with the pipet tip or sonicator for complete washing. Repeat, spin again, and discard the final supernatant.
6. Allow residual acetone to air dry until the tube is odorless (∼5 min, RT). TIP: Careful not to overdry or the pellet will be difficult to resuspend.

8.2.3.2. Solubilization and digestion.

7. Solubilize the pellet in step 1 digestion buffer. TIP: Do not heat samples in urea. If the pellet is difficult to solubilize, sonicate in an ice water bath instead.
8. Estimate protein concentration using a Bradford assay.
9. Reduce disulfide bonds by adding 1 µg DTT per 50 µg protein, incubate at RT, 30 min.
10. Alkylate cysteine residues by adding 5 µg IAA per 50 µg protein, incubate at RT, 20 min. TIP: IAA is light-sensitive – keep the reaction and reagents in the dark.
11. Digest protein by adding 1 µg lys-C per 50 µg protein, incubate at RT, 3 h, dark.
12. Dilute the mixture with six volumes of step 2 digestion buffer and add 1 µg trypsin per 50 µg protein, incubate at RT at least 4 h (up to overnight), dark.
13. When complete, acidify the peptide digest to pH <2.0 using 20% formic acid diluted in water. Check pH by dispensing a small amount (∼0.2 µl) onto a pH test strip.

8.2.3.3. Peptide desalting and resuspension.

14. Prepare desalting "columns" (Rappsilber et al., 2003) by punching out small disks of C18 Empore filter using a 17 G flat-tipped syringe and ejecting the disks into P200 pipette tips. Ensure that the disk is securely wedged in the bottom of the tip. Stack bulk C18 powder above the disk by suspending the powder in methanol and pipetting it into the tip. Elute the methanol either by centrifugation or by pressurizing the tip with a syringe barrel. The C18 stack should be ∼1 cm tall and can desalt a sample containing ∼50 µg of peptides, although capacity may vary depending on the type of bulk C18 material used.
15. Wash the column with 200 µl Buffer B. Discard eluate. If any C18 columns have visible channels or dead volume after this stage, do not use the column.
16. Condition the column with 200 µl Buffer A. Discard eluate.
17. Load the column with the peptide digest sample. Discard eluate.
18. Wash the column at least twice with 200 µl Buffer A. Discard eluate.
19. Elute peptides into a clean tube with 200 µl elution buffer. Repeat for complete elution (expect one elution to yield ∼90% of peptides).
20. Evaporate the sample in a vacuum centrifuge (RT) until dry.
21. Dissolve peptides in resuspension buffer and quantify by nanodrop. Equalize the sample concentrations. TIP: If an absorbance peak is observed at 240 nm, this likely indicates residual urea/thiourea contamination. The sample may need to be desalted again
22. Centrifuge diluted samples at 16,000 g (10 min) to remove potential particulates from solution. Transfer the supernatants in randomized order to a 96-well autosampler plate for LC-MS/MS.

8.3. Liquid chromatography and mass spectrometry

There are many types of columns, chromatography systems, and mass spectrometers that can be used for proteomics, precluding an universal standard method. Here we offer four main points to consider when deciding on a specific approach.

1. The amount of sample to inject depends on the sample complexity, chromatography resolution, and instrument sensitivity, but it is normally in the range of 0.1 – 1 µg (lower complexity samples will require less material to inject).
2. If a sample has high complexity but is dominated by a relatively small number of highly abundant peptides, such as what is routinely observed in mammalian plasma (Hortin & Sviridov, 2010) or honey bee ejaculates (McAfee

et al., 2020), the sample may benefit from orthogonal fractionation upstream of LC-MS/MS. That is, the peptides should be fractionated based on different chemical or physical properties than utilized in the downstream LC (which normally separates peptides based on hydrophobicity) in order to reduce ion suppression by the highly abundant species. Examples of orthogonal or semi-orthogonal fractionation techniques are high-pH reverse-phase fractionation (Batth et al., 2014), strong cation exchange (Edelmann, 2011), size exclusion chromatography (Kristensen et al., 2012), and isoelectric trapping (Cologna et al., 2010).

3. The final stage of chromatography, which is online-coupled to the mass spectrometer, is nearly always C18 reverse phase run under acidic conditions. Different chromatography systems have different constraints, but in general, the length of the chromatography gradient trades off peak intensity with peak separation: longer gradients offer better separation, but eluent peaks are broader with lower intensity, which may affect spectrum quality. Gradients typically ramp from 2 to 80% acetonitrile over the course of 45–180 minutes. Refer to vendor recommendations for information on column selection and instrument settings.
4. There are two main data acquisition approaches for shot-gun proteomics: data-dependent acquisition (DDA) and data-independent acquisition (DIA) (Doerr, 2014). DIA is growing in popularity, as it is thought to yield more accurate, more reproducible, and higher coverage proteomic data (Barkovits et al., 2020), though this point is debated (Fernández-Costa et al., 2020). DIA has, at the time of writing, not yet been applied to honey bees, and the approach has unique database considerations if a spectral library is used (Pino et al., 2020). The sample preparation methods covered here can be used for both DDA and DIA, but downstream data processing may differ. This is a technique to watch as it becomes more popular and widespread.

8.4. Proteomics data processing

8.4.1. Software recommendations

Raw mass spectrometry data must be "searched" in order to derive biological meaning from it. That is, a computer algorithm matches mass spectra to peptide sequences, deconvolutes the pool of identified peptides into the most parsimonious set of proteins that must be present to explain all those peptides, translates peptide intensity data into quantitative protein data, and controls false discovery rates. While there are many software options available to perform these tasks (reviewed elsewhere (Verheggen et al., 2020)), we recommend MaxQuant (Cox & Mann, 2008) owing to its (i) high regard in the field, (ii) robust label-free quantification (LFQ) algorithm, (iii) delayed normalization feature to accommodate fractionated samples, (iv) continual feature upgrades, (v) capability of handling DIA and DDA data, and (vi) lack of associated cost (please see the latest information from the annual **MaxQuant** Summer School for upcoming tutorials: https://maxquant.org/summer_school/). DIA-NN is also an excellent search tool which is especially well-suited for processing DIA data (Demichev et al., 2020). **MaxQuant** and **DIA-NN** are only compatible with Windows and Linux machines. Theoretically, spectra can also be sequenced de novo (without an existing protein database) using other software, such as **PEAKS** (Ma et al., 2003); however, this is only viable for very high-quality spectra and the resulting proteome coverage is therefore exceedingly low.

8.4.1.1. MaxQuant and DIA-NN search parameters.

The default search parameters within **MaxQuant** and **DIA-NN** are generally appropriate for most shot-gun proteomics analyses, but some options should be considered (and see Sinitcyn et al. (2021) for handling DIA data in **MaxQuant**). In particular, "match between runs" or "MBR" is an option that increases sensitivity by borrowing peptide identification information across samples. For example, if a spectrum is confidently matched to a peptide in sample 1 but not sample 2, sample 2 is re-inspected for likely features of that spectrum, and, if found, it can receive the same peptide assignment. This approach assumes that if a peptide is confidently identified in one sample, it has a high likelihood of being present in other, similar samples, and the spectrum quality threshold for that peptide matching can be reasonably lowered. We recommend enabling this option in both **MaxQuant** and **DIA-NN** to reduce the frequency of missing data. In **DIA-NN**, "unrelated runs" should also be checked if samples represent independent replicates, and the "Protein inference" option should be set to "Protein names (from FASTA)."

8.4.1.2. Choosing an appropriate protein database.

For typical shot-gun proteomics experiments, the data processing software requires at least two inputs: the raw data files and a database of proteins to which it can compare spectra. It is very important to choose an appropriate protein database; failure to do so can result in flawed data with an unacceptable level of false positive or false negative errors. One example of this happening in the literature includes

a published paper claiming to discover a link between invertebrate iridescent virus-6, detected in honey bee proteomics samples, and colony collapse disorder (CCD) (Bromenshenk et al., 2010). The database used to search the mass spectrometry data included only viral protein sequences and no host (honey bee) proteins, despite host proteins composing the majority of the sample.

This means that, since spectrum matching is a probabilistic task, it is possible for spectra from host peptides to match to viral peptides if those are the most likely assignments within the constraints of the protein database supplied. Indeed, that is exactly what happened, leading to incorrect peptide assignments, dramatically skewed false discoveries, and ultimately flawed conclusions. When the host proteins were included in the search database, spectra that previously matched to iridescent virus-6 actually had far higher scoring matches to host peptides, indicating that the virus was unlikely to have actually existed in the sample (Foster, 2011; Tokarz et al., 2011), let alone cause CCD.

Since genome builds, as well as gene and protein annotation databases, are continually upgraded, the most up-to-date reference proteome should be used. Furthermore, and following the above discussion, the database should contain all sequences with a reasonable probability of being found in the sample. For honey bees, this means that, in addition to honey bee protein sequences, honey bee virus sequences should be included in the protein database (FASTA file) for virtually all sample types, given the high incidence of asymptomatic infections (Grozinger & Flenniken, 2019). Nosema spp., chalkbrood (*Ascosphaera apis*), European foulbrood (*Melissococcus plutonius*), American foulbrood (*Paenibacillus larvae*), or any other likely pathogen or colonizing microbe may be added as well, if applicable. We recommend obtaining FASTA files from Uniprot due to the ease of subsequently incorporating gene ontology (GO) information during data analysis. We also recommend including protein sets for the core gut bacteria (Motta & Moran, 2024) when bee abdomens form part of the sample (see Section 10).

8.4.1.3 Statistical analysis. When finished searching, **MaxQuant** will output a series of tables, including one named ProteinGroups.txt, which contains the protein quantitation information with the dominant members of the protein groups and LFQ intensities. The equivalent output from **DIA-NN** is report.pg_matrix.tsv file. The matrix of protein names (rows) and LFQ intensities (columns) is used for subsequent differential expression analyses. Any proteins indicated as reverse hits, potential contaminants, or those only identified by site are undesirable and typically excluded. **MaxQuant's** companion program, **Perseus** (Tyanova et al., 2016), can be used for basic statistical tests and figure generation; however, most R packages originally intended for microarray or RNA-seq data analysis (e.g., **limma** (Ritchie et al., 2015)) are also appropriate for proteomics data and offer more flexibility. For users who are new to proteomics analysis, we recommend using **Perseus**, since it is an user-friendly platform developed specifically for proteomics, and is accompanied by detailed step-by-step tutorials (http://www.coxdocs.org/doku.php?id=perseus:user:use_cases:interactions) and lectures (http://www.coxdocs.org/doku.php?id=perseus:user:tutorials). The tutorial "label-free interaction data" provides a detailed guide to data preparation (loading, filtering, transforming, etc.), quality control, statistical analyses, and visualization. Currently, **Perseus** is only compatible with Windows.

Once differential expression analysis is complete, the results may be used for gene ontology (GO) term enrichment tests similar to what might be conducted for microarray or RNA-seq data. While a multitude of suitable tools exist for such analyses, reviewed in Laukens et al. (2015), we recommend ErmineJ (Gillis et al., 2010; Lee et al., 2005) for its flexibility, simplicity, and capability for accounting for both multiple hypothesis testing and protein multifunctionality when determining enrichment significance.

8.5. Applications and limitations

Shot-gun proteomics can be used to investigate anything from responses to pesticides, pathogens, nutritional stress, aging, and an endless array of other conditions (Arad et al., 2024). Proteomics has even been used to discover specific protein markers suitable for guiding selective breeding for Varroa resistance mechanisms (Guarna et al., 2017). While LFQ proteomics is best suited to compare conditions across which the majority of proteins can be assumed to be expressed at the same abundance, with a smaller fraction of the proteome changing in response to a stimulus, meaningful results can be obtained from experiments with more dramatic proteomic shifts, such as between castes and across developmental stages. LFQ has been demonstrated to achieve accurate relative quantification even when approximately one-third of the proteome is changing in abundance (Cox et al., 2014).

Although proteomics is an increasingly powerful technique, interpretation of the results can be challenging owing to limited functional annotation of the honey bee proteome (Elsik et al., 2018). Each protein can have one or more biological functions,

which are associated with unique GO terms (Ashburner et al., 2000), to help derive biological meaning from the hundreds or thousands of proteins that are often differentially expressed in proteomics experiments. According to Hymenopteramine, a database of genomic resources for hymenopterans, only 7,929 out of 15,314 sequences (52%) in the honey bee official gene set (v3.2) were linked to GO terms as of 2018 (Elsik et al., 2018), meaning that the remaining sequences have poorly characterized functions. This figure has since increased (Walsh et al., 2022), but many uncharacterized genes remain. Our limited understanding of honey bee gene and protein functions means that high-throughput datasets can be difficult to interpret, as we are blind to the roles of a large fraction of the very targets we are analyzing.

Because shot-gun proteomics requires protein digestion into peptides, and peptide sequences can be shared between different proteins, it is often difficult to say definitively to which protein the peptides belong. MaxQuant deals with this problem by reporting "protein groups," which offer the most parsimonious explanatory proteins likely to be present in the sample, rather than individual proteins. However, this also complicates GO term assignment: Since multiple different proteins can be listed in a single protein group, which GO terms should the protein group be given? A simple heuristic, though imperfect, is to assign a protein group with the GO terms associated with its leading protein. Since both GO terms and protein groups are defined based on sequence similarities, it is a reasonable assumption that proteins within a group will share GO terms. Alternatively, though more laboriously, GO terms associated with all proteins in the group can be linked.

Unfortunately, for those proteins which are poorly characterized, it is difficult to generate functional information without a high-throughput way to generate gene knock-out (a gene is deleted or rendered nonfunctional) or knock-in (a gene is inserted) mutant organisms. Organisms such as Drosophila melanogaster and Mus musculus have benefitted from decades of detailed genetic and biochemical research into specific genes and proteins, but this is only recently possible for honey bees and is still far from routine (Kohno & Kubo, 2019). While much information can be borrowed from what is known about homologous proteins in other species, honey bees diverged from flies about 300 million years ago (Honeybee Genome Sequencing Consortium, 2006) and therefore have experienced considerable sequence divergence. Until we know more about the functions of all honey bee proteins, we will not be able to interpret high-throughput differential expression data to its full potential.

9. Metabolomics

9.1. Introduction

Metabolomics is the study of small molecules, metabolites, and biological intermediate substrates. This omics tool has become very popular within the last decade, including among entomologists (Snart et al., 2015), and applications to understand the biology of honey bees (Ardalani et al., 2021; Broadrup et al., 2019; Chandrasekaran et al., 2015; Chang et al., 2022; Chen, Wang et al., 2021; Du Rand et al., 2017; Jousse et al., 2020; Klupczynska et al., 2020; Li et al., 2020; Ma et al., 2024; Paten et al., 2022; Pratavieira et al., 2020; Rand et al., 2015; Ricigliano et al., 2022; Rothman et al., 2019; Shi et al., 2018; Wang, Habermehl, et al., 2022; Wu et al., 2017; 2024; Xu et al., 2024; Zhao et al., 2020; Zhong et al., 2024), the relationship with their symbionts (Kešnerová et al., 2017; Quinn et al., 2024; Zhang, Mu, Cao, et al., 2022; Zhang, Mu, Shi, et al., 2022; Zheng et al., 2017), and characteristics of colony products (Arathi et al., 2018; Baky et al., 2023; Chakrabarti et al., 2019; Guo et al., 2020; Koulis et al., 2021; Li et al., 2019; Milone et al., 2021; Qi et al., 2023; Sun et al., 2021; Virgiliou et al., 2020; Wang, Li, et al., 2022; Wilson et al., 2013; Yan et al., 2024; Yusoff et al., 2022) are expanding (see Jung (2023) for a brief review). The method allows small molecules to be characterized in a biological system, and is an important complement to more established omics methods, such as genomics, transcriptomics and proteomics. In fact, metabolomics is often considered as the final piece of the omics puzzle (Veenstra, 2012). Metabolomic investigations can be conducted using nuclear magnetic resonance (NMR), capillary electrophoresis mass spectrometry (CE-MS), gas chromatography mass spectrometry (GC-MS) or liquid chromatography tandem mass spectrometry (LC-MSMS) (reviewed in Munjal et al. (2022)). LC-MSMS is the most commonly used technique and is the one we focus on here.

Unlike the other omics tools, where species-specific sequence libraries are required, one advantage of metabolomics is that a general library of small molecule fragmentation patterns has been established. Fragmentation patterns and other characteristics (mass, isotope ratio, and retention time) determine the metabolite being identified, rather than a specific sequence of the organism. Metabolomics is thus a widely applicable tool, especially for honey bees, where researchers are trying to assess the biological processes of development or the physiological impacts of various stressors (pesticides, malnutrition, mites, etc.) or other stimuli.

The following general method has been successfully applied to analyze honey bee pollen (Chakrabarti et al., 2019) and royal jelly (Milone et al., 2021), but also works well for conducting metabolite detection in honey bee tissues, where the method has enabled identification of 251 high-confidence metabolites from whole honey bees (Chakrabarti et al., manuscript in preparation). However, as there are many sample types and desired compound classes of potential interest, variations of this protocol may work better for certain applications. For example, while methanol/water extractions are commonly conducted (e.g., Paten et al., 2022; Chang et al., 2022; Xu et al., 2024), extraction solvents composed of a mix of acetonitrile, methanol, and water (e.g., Chen, Wang, et al., 2021; Ma et al., 2024), acetonitrile and methanol (e.g., Liu et al., 2023; Wu et al., 2024), or methanol and chloroform (e.g., Ricigliano et al. (2022)) have also been used for honey bee tissues. The choice of extraction solvent depends on the polarity of the desired metabolites, and additional extraction techniques developed for mammalian tissues (Sitnikov et al., 2016) are likely also applicable to honey bees. For example, in a preprint report, (McAfee et al., 2024) recently applied a two-phase extraction technique (using an initial methanol/water extraction followed by addition of methylated tert butyl ether) developed by Chen et al. (2013) for mouse liver samples for parallel analysis of the metabolome, lipidome, and pheromone profiles from queen honey bee heads. The following protocol serves as a starting point for researchers interested in conducting metabolomics, with the knowledge that there are many possible variations of the general technique, particularly with respect to the extraction solvent and liquid chromatography solvents.

The method described here is intended for a semi-quantitative (i.e., relative quantitation, which is suitable for differential abundance testing but does not provide information on absolute quantities), untargeted metabolomics approach. These sample preparation guidelines are also applicable to absolute quantitation, provided that the user includes the additional step of creating standard curves for each analyte. However, as untargeted semi-quantitative analysis is the most widely used method, that is the application we focus on for both sample preparation and data analysis methods. Interested readers should familiarize themselves with existing overviews of sample preparation and data acquisition (Broadhurst et al., 2018; David & Rostkowski, 2020; Defossez et al., 2023; Munjal et al., 2022; Rampler et al., 2021), statistical analysis (Bartel et al., 2013; Chen et al., 2022; Xi et al., 2014), and reporting (Alseekh et al., 2021; Sumner et al., 2007) described elsewhere.

9.2. Sample preparation for metabolomics

9.2.1. Considerations

9.2.1.1 General

- Metabolomics experiments may be either focused on discovery (untargeted analysis; e.g., discovery of metabolome changes in response to stimuli) or monitoring changes in or presence of an a priori defined set of compounds (targeted analysis). This protocol may need to be adjusted if the analytes of interest are better extracted in another solvent or better separated with another LC gradient.
- Untargeted metabolomics analysis is normally conducted using high-resolution instruments such as Q-TOF or Orbitrap mass spectrometers while targeted metabolomics may be performed with low resolution, highly selective mass spectrometers such as triple quadrupoles (QQQs). Ensure the instrumentation available meets the needs of your experiment and sample complexity
- Mass spectrometer sensitivity may drift between runs. Running QCs every few samples is thus useful for data normalization. Constitute quality control (QC) samples by mixing equal volumes from all samples.
- Data acquisition can be performed in both positive and negative ion mode, and small molecules may be more amenable to one or the other depending on their structure and functional groups. Though time consuming, analyzing samples in both negative and positive ionization modes captures a wider spectrum of metabolites.
- Aim to inject 1–10 μl of sample into the column. Before running a batch of new samples, check the concentrations by running a diluted test sample. This will help gauge the amount to inject to avoid overloading or underloading the instrument.
- Adding a generic internal standard during sample preparation can help the user compare data between sample runs and facilitate relative quantitation. However, for absolute quantitation, either isotopically-labeled internal standards matching the compound of interest or external standard curves of the unlabelled compound are required.

9.2.1.2. Sample handling

- Store samples at −80 °C
- Sample processing is somewhat dependent on the type of metabolite to be extracted and identified. Generally, methanol:water extractions work well, but the extraction solvent may need to be adjusted to best capture specific classes of compounds (see Section 9.1).
- The step at which an internal standard is added may vary. Adding the standard before

homogenization accounts for variation in extraction efficiency, while adding the standard immediately ahead of injection will ensure the same amount is present in each sample. If desired, two sets of internal standards may be added at different points in the sample processing protocol.

- Blank samples should be prepared in parallel with real samples in order to determine what compounds are contaminants introduced from solvents, plastics, and the environment. These blank samples should be analyzed at regular intervals during data acquisition and it is good practice to either subtract the average intensities of compounds in the blank samples from those in real samples, or to only retain sample compounds present at an abundance above a threshold factor over the blanks (e.g., 5-fold higher in real samples than blanks)
- Solvent-only injections may also be performed at regular intervals in order to assess sample carryover between injections as well as potential system contamination. These injections are distinct from blank samples in that they are not derived from parallel processing of real samples; instead, they are comprised of pure reconstitution buffer
- LC-MS/MS instruments with auto-samplers require HPLC vials (~2 mL). For small volume samples, a 300 µl vial insert may be required.

9.2.1.3. Reagent handling

- All reagents and solvents should be mass spectrometry grade. Care must be taken to avoid any impurities in the reagents and chemicals used.
- Prepare fresh buffers and solutions.

9.2.2. Materials

- Large equipment: homogenizer or bead beater, centrifuge, speed vacuum concentrator, -80 °C and −20 °C freezers, a vortex, and liquid chromatography system coupled to a mass spectrometer (for example, a high-resolution system such as a Nexera LC30 UPLC (Shimadzu) coupled to a quadrupole-time-of-flight mass spectrometer (TripleTOF 5600, AB SCIEX), but see Section 9.3)
- In-house library of metabolites (such as the IROA Mass Spectrometry Metabolite Library of Standards)
- Mass spectrometry grade methanol, water, formic acid, and acetonitrile
- Homogenization tubes and beads

9.2.3. Metabolomics methods

9.2.3.1 Sample homogenization

1. Homogenize 50 mg of sample in 0.5 ml of methanol and water solution (80:20 *v/v*) in a homogenizer (for example, a QIAGEN TissueLyser or Precellys 24 tissue homogenizer). Use tubes and tips without color. TIP: Including BHT (butylated hydroxytoluene) at 0.01% in the extraction solvent protects compounds from oxidation
2. Homogenize until the tissues are broken down into minute particles to maximize extraction.

9.2.3.2. Extraction

3. Incubate samples for at least 1 hour at −20 °C to allow metabolites to be extracted and proteins to precipitate.
4. Centrifuge at 10,000 g for 10 min (4 °C) to pellet the precipitate containing proteins and other cellular debris.
5. Transfer 400 µl of the supernatant to a clean 2 ml tube and evaporate to dryness in a speed vacuum concentrator.
6. Reconstitute dry extracts in 200 µl of acetonitrile:-water solution (1:1 *v/v*).
7. Vortex the tube for 30 s and centrifuge for 10 min at 10,000 g (4 °C).
8. Collect the supernatants and transfer them to HPLC vials.
9. If not immediately conducting mass spectrometry, store the samples at −80 °C until analysis. Re-centrifuge the samples before LC-MS/MS to remove any precipitates.

9.3. Chromatography and mass spectrometry

Similar to proteomics (see Section 8), there are numerous options for chromatography and mass spectrometer instruments and acquisition methods. Additional considerations for chromatography and mass spectrometry are as follows:

1. The methods described here have been used successfully with a Nexera LC30 UPLC (Shimadzu), coupled to a quadrupole-time-of-flight mass spectrometer (ABSCIEX TripleTOF 5600); however, other combinations are possible. High-resolution mass spectrometers require occasional calibration. Time of flight mass spectrometers (TOFs), for example, require calibration every 2 to 3 hours to retain <3 ppm mass accuracy.
2. The HPLC column is used to separate metabolites and other small molecules based on polarity (Figure 19). The LC-MS/MS approach described here has been used with an Inertsil Phenyl-3 stationary phase column (such as 150 × 4.6 mm, 5 µm by GL Sciences), but other columns may be better suited for analyzing different kinds of metabolites. Common columns

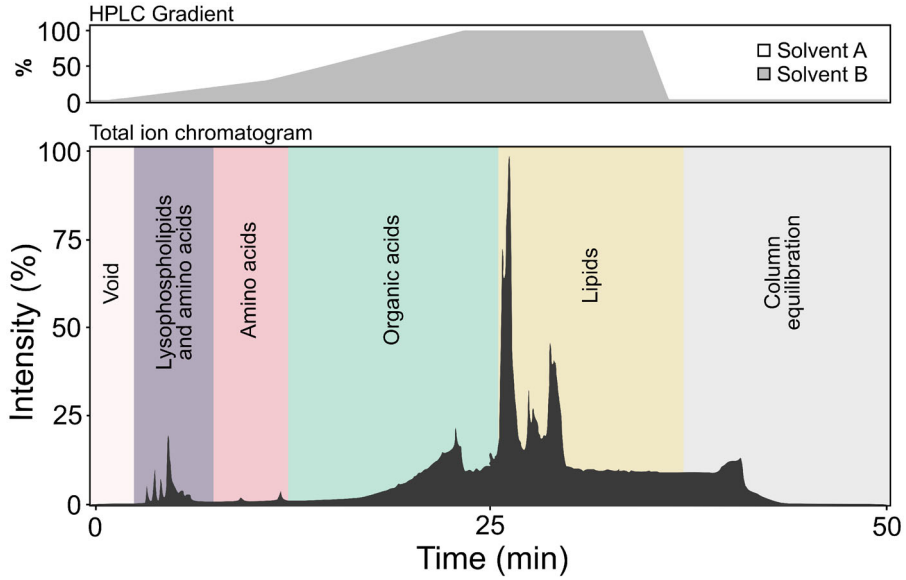

Figure 19. Example of the chromatographic profiles TIC (total ion chromatogram). Typical compound classes; amino acids, lipids and organic acids are identified in the different regions during the HPLC gradient.

include BEH C18 and ACQUITY HSS T3 (both manufactured by Waters) because they are suitable for analytes with a broad range of molecular weights and polarity.

3. A guard column can be used to avoid unrecoverable column contamination. In case of contamination or sample impurities, the guard column can be discarded and replaced, thereby retaining the functionality of the HPLC column and extending its lifespan.
4. For specialized small molecule analyses, an entirely different extraction and analysis method may be needed. Queen mandibular pheromone, for example, is typically extracted in ether, derivatized, and analyzed by GC-MS), rather than LC-MSMS, although a recent report does describe successful utilization of LC-MSMS for this purpose (McAfee et al., 2024). Some testing may be required to see which method works best for your target analytes. In addition, small and/or highly polar metabolites may not be compatible with reverse-phase chromatography; such compounds may require specialized chromatography techniques (e.g., hydrophilic interaction liquid chromatography, or HILIC; (Buszewski & Noga, 2012) or derivatization and analysis by GC-MS (De Souza, 2013).
5. Data is normally acquired in data-dependent acquisition mode (DDA; also known as information dependent acquisition, or IDA). Survey scans are acquired followed by a specified number of MS/MS spectra in a given scan. This results in a "feature," i.e., a mass, retention time, isotope ratio, and an MS/MS fragmentation pattern (Figure 20). In DDA mode, parent ion peaks are selected for fragmentation based on their intensities (ions producing the most intense peaks are selected). The number of ions that can be fragmented is finite and depends on the speed of the instrumentation. Data-independent acquisition (DIA) is also available for metabolomics analysis, and although it results in lower spectrum quality than DDA, DIA appears to be more precise and has better fragmentation spectrum coverage (Guo & Huan, 2020).
6. Sample injections should always be randomized. In addition, blanks and QCs should be run at the beginning, during, and at the end of injection sequences to help correct for drifting signal intensities. This is especially important for large batch metabolomics.
7. Gradients for metabolite elution may vary. This is experimentally derived to separate and optimize metabolite identification. For honey bee metabolomics, the following gradient has been successfully used (solvent A: 100% water containing 0.1% formic acid; solvent B: 100% methanol containing 0.1% formic acid): 0.0 min @ 5% Solvent B; 1 min @ 5% solvent B; 11 min @ 30% solvent B; 23 min @ 100% solvent B: 35 Min @ 100% solvent B; 37 min @ 5% solvent B; 50 minutes stop chromatography. While gradients using methanol are common (e.g., Liu et al., 2023, McAfee et al., 2024), it may also be appropriate to use acetonitrile (e.g., Shi et al., 2018; Paten et al., 2022; Wu et al., 2024; Zhong et al., 2024) and additives such as ammonium formate (e.g., Xu et al., 2024), or ammonium acetate and ammonium hydroxide (e.g., Ma et al., 2024), depending on the chemical properties of the analytes to be separated. Refer to the vendor instructions when choosing the column

and guard column, flow rate, solvent composition, and column cleaning tips. A GL Sciences Phenyl 3 column (4.6 × 150mm, 5 uM particle size) was used for this metabolomics profile.

9.4. Metabolomics data processing

The essential steps of metabolomics data processing that are required ahead of statistical analysis are peak picking (selecting chromatographic peaks), alignment (correcting for shifts in retention times), deconvolution (separating composite spectra from co-eluting analytes), integration (calculating peak areas), normalization (correcting for variation in sample injection amounts), and database querying/searching (annotating peaks with compound identities). The final result of these processing steps is a data matrix of samples and high-confidence compounds with their intensities that can be used for statistical analysis (differential abundance testing, pathway enrichments, etc.).

These informatic tasks can only be achieved with the help of specialized software (see Chen et al. (2022) for an outline of workflows and software options), which may include either commercial vendor software, free software available online, or a combination of both. In some cases, all essential data processing steps can be achieved within one software platform, whereas others may require a combination of tools (see below for examples). Regardless of the workflow employed, users should be familiar with the minimum reporting criteria outlined by the Metabolomics Standards Initiative before commencing (Spicer et al., 2017; Sumner et al., 2007).

Examples of paid software include Peakview (AB SCIEX) for peak visualization and compound identification, and MultiQuant (AB SCIEX) for feature integration, among other instrument-specific options. Progenesis QI (Waters, Nonlinear dynamics) is another popular paid software that can achieve all aspects of raw data processing when accompanied by the METLIN (discussed further below) plugin for database querying. Reifycs is another cross-instrument paid software option which supports various data formats and negates the need to purchase dedicated software for each mass spectrometry instrument.

Freely available software is also highly regarded in the field and can be used to process data from many types of instruments, provided the raw spectrum data files can be converted to the required

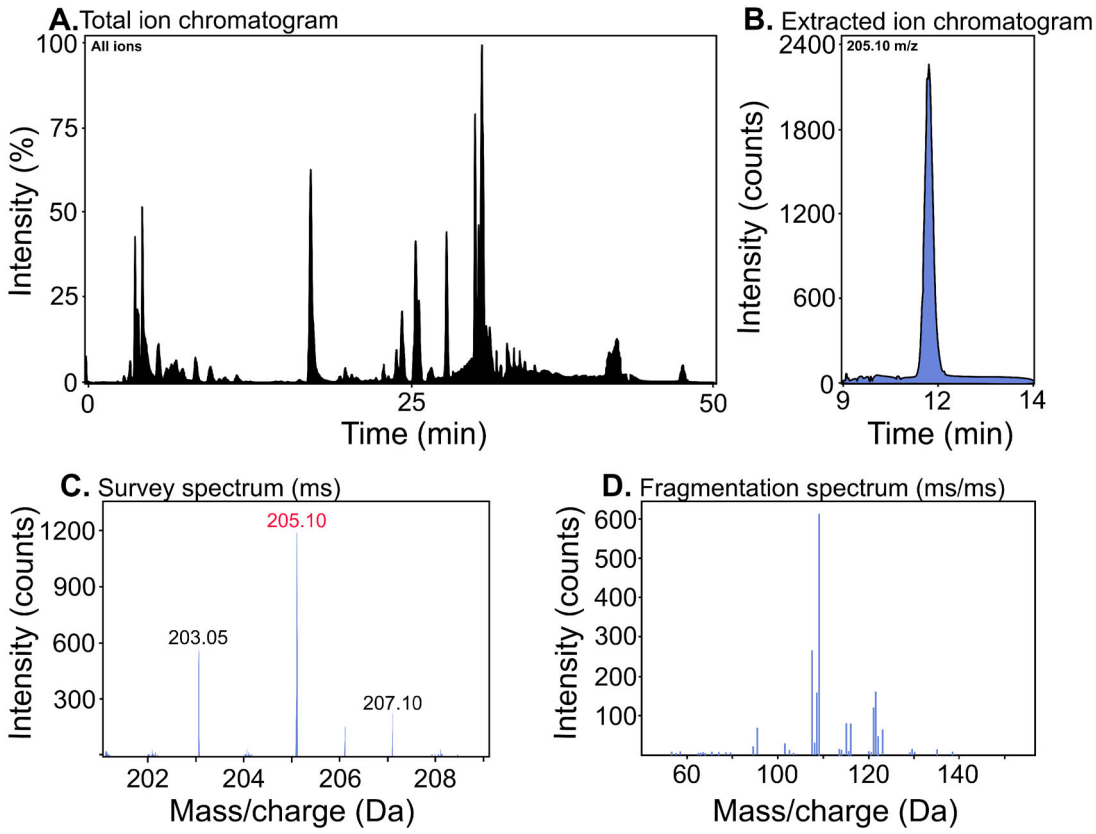

Figure 20. Description of an IDA (information dependent acquisition) feature. (A) A total ion chromatogram of all ions detected in the sample. Each ion (feature) has a corresponding (B) Extracted ion chromatogram (in this case, for the 205.10 parent ion), (C) Survey spectrum scan containing the ion at a given retention time, and (D) Fragmentation spectrum for the given parent ion and retention time.

input format. For example, MetaboAnalyst is one of the most widely used metabolomics data processing tools and can be used for several raw data types, once converted from the vendor data format to NetCDF, mzXML, or mzDATA format. The newest release of MetaboAnalyst (v. 6.0) includes updated processing and peak annotation modules for fragmentation (MS/MS) spectra and more sophisticated downstream statistical analysis modules (Pang et al., 2024). MSDial is another popular tool for raw data processing, which requires data in analysis base file (ABF) format (Tsugawa et al., 2015). The task of converting data files to the formats required by different software can be achieved using open-source tools available through ProteoWizard or tools such as Reifycs Abf (for conversion to ABF format; available at https://www.reifycs.com/AbfConverter/index.html).

The ultimate goal of metabolomics is to identify and quantify small molecule metabolites in a biological system, so, clearly, the task of assigning spectra to correct compound identities is critical. After peak picking, alignment, deconvolution, integration, and normalization, metabolites are tentatively identified by matching their masses, fragmentation patterns, isotope distributions, and retention times (Figure 21), to corresponding data from an in-house chemical library, *in silico* library, or both. An in-house chemical library is a purchased set of metabolite standards (e.g., the IROA library) that are used to acquire reference data using the same conditions as the metabolomics HPLC gradient, whereas an in silico library is a database of spectra and compound identities (which should be derived from the same instrumentation used to acquire the experimental data). Examples of online in silico resources include METLIN, NIST17, MassBank Europe (mass spectral database of Europe) and MoNA MS/MS libraries (MassBank of North America) (Ardalani et al., 2021). The METLIN database used to be publicly available (Guijas et al., 2018), but the updated

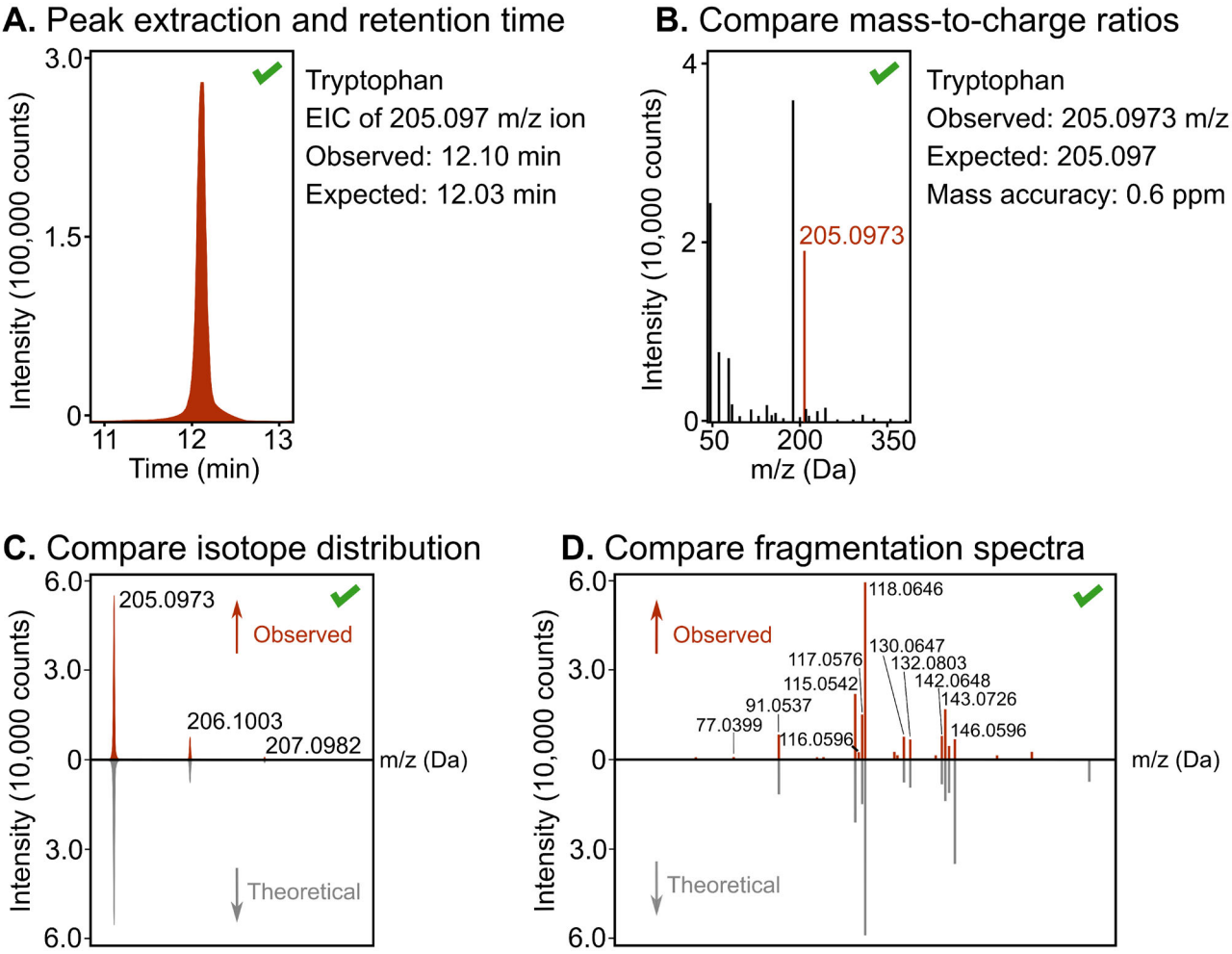

Figure 21. Example of methodology for tentative metabolite identification. Tryptophan is used in this example. (A) The extracted ion chromatogram is compared with an established HPLC method library with known retention times for the metabolite of interest. (B) The survey spectrum to compare mass measurements. Mass error should be less than 5 ppm. In this example the mass error is 0.6 ppm. (C) Isotope pattern. The theoretical isotope distribution for tryptophan is compared to the observed isotope ratio. (D) Library MS/MS spectrum matching. Observed MS/MS is compared to the IROA library MS/MS spectrum for tryptophan.

version (XCMS-METLIN) is now behind a paywall. NIST17 is also a paid resource, with mass spectral libraries as well as various software tools that are needed to analyze the datasets. MoNA is a free database currently housing over two million mass spectral records from multiple data sources, such as experimental libraries with established datasets, in silico libraries, as well as information sourced from general user contributions. Regardless of the software or libraries used, it is essential to manually check each feature's integration and qualifying attributes before including them in the list of tentatively identified metabolites, otherwise the risk of misassignment is high.

Once the data are extracted and in ".xls" or ".csv" format, a wide range of statistics and visualization methods can be used, such as producing heat maps, hierarchical clustering, principal component analysis, and correlation matrices, or performing comparisons using t-tests or ANOVAs (with multiple hypothesis testing corrections), etc. These statistical approaches are essentially the same as for other high-throughput biological data, such as proteomics or RNA-seq. To visualize data (as, e.g., PCA plots, heat maps, dendrograms, etc.), analytical software such as MarkerView (AB SCIEX) or Progenesis QI can be used. As mentioned earlier, the freely available software MetaboAnalyst, which enables a user to directly upload mass spectrometry files to the online platform, is especially appealing with its updated statistical analysis module (it can now handle more complex experimental designs) and new module for estimating causal relationships between metabolites and phenotypes (Pang et al., 2024). This software can also compute pathway analysis, enrichment analysis, biomarker analysis, network analysis, joint pathway analysis, and has an associated package for use in R (MetaboAnalystR; (Chong & Xia, 2018)). Please see Chen et al. (2022) for additional statistical analysis workflows and methods.

9.5. Metabolomics applications and limitations

Metabolomics is gaining popularity due to the vast spectrum of metabolites that can be identified and compared between samples. Using mass spectrometry for metabolomics has many advantages over traditional biochemical assays, such as high sensitivity and the ability to detect a large number of metabolites and small molecules from very small sample sizes (Veenstra, 2012). In addition, with the availability of better databases, identification of small molecules has become easier. But despite the effort that goes into sample preparation and analysis, metabolomic identifications are only tentative detections until retention times and fragmentation patterns are confirmed with high-purity analytical standards, whether part of an in-house library or purchased separately. When this confirmation is achieved, such compounds may rise to the level of being "identified" (i.e., level 1 annotation, as opposed to levels 2, 3, and 4, corresponding to "putatively annotated compounds," "putatively annotated compound classes," and "unknown compounds," respectively (Sumner et al., 2007).

Features must always be compared with a MS/MS spectral library (e.g., METLIN) when compounds are not an exact match with those within the in-house library. In this case, a high-purity standard must be purchased to confirm a tentative assignment. Furthermore, for absolute quantitation of targeted metabolomics, isotopically labeled and unlabeled standards are required for exact quantifications of the targeted small molecules. While labeled standards are used as internal standards, unlabeled standards are used for creating standard curves for absolute quantification. All these analytical standards can add a substantial cost to an already costly method, and labeled standards are not available for every compound, but these steps are essential to confidently identify and quantify small molecules.

If metabolomics is conducted in the absence of an in-house chemical library, the user must rely on generic spectral matching using digital libraries for putative compound assignments, and reliability of such matches is limited without time consuming manual assessments of annotations and subsequent verification using analytical standards. However, reliable results can still be produced using open source software tools not made by a specific instrument vendor, particularly by integrating complementary digital reference libraries, and confirming identifications with pure analytical standards.

Metabolomics is a powerful tool and gives us a snapshot into a wide spectrum of biological molecules in honey bees. It is helpful for understanding honey bee developmental physiology or adaptive molecular responses to various stressors. With ongoing small molecule discoveries and inclusion of these species in spectral libraries, metabolomics is emerging as the new power tool in modern omics analyses. Excitingly, the closely related field of lipidomics is also emerging in honey bees (Morfin et al., 2022), and co-extraction of metabolites, lipids, and pheromones has been performed on queens (McAfee et al., 2024). The diversity of small molecules that can be analyzed by LC-MSMS and integration with pheromone analysis will likely enable many new insights into honey bee physiology in the future.

10. Microbiome analysis

10.1. Introduction

The increased concern for honey bee health has led to a collective effort by researchers to unravel the major contributors to honey bee fitness. Increasingly, the microbiome – the suite of microorganisms, from their genes to metabolites – is thought to play a major role in honey bee health. Nine main bacterial taxa compose the gut microbiome of honey bee workers, and among them, Snodgrassella, Gilliamella, Bombilactobacillus, Lactobacillus nr. melliventris, and Bifidobacterium form what is known as the core gut microbiome (Zheng et al., 2018). Thus far, we know these gut-associated bacteria contribute to the honey bee's nutrition (Engel & Moran, 2013; Kešnerová et al., 2017), immune system (Kwong, Mancenido, et al., 2017), detoxification of xenobiotics (Wu, Zheng, et al., 2020), response against pathogens (Raymann et al., 2018), colony structure through nestmate recognition (Vernier et al., 2020), and several other roles that ultimately impact colony fitness (Figure 22(A)).

Though the interconnectedness between honey bee gut microbiota and honey bee health is undeniable, the interconnectedness between honey bee health and the microbiome of hive interiors should also be taken into account. Honey bees are highly eusocial insects, meaning the entire colony – individuals, environments, and associated microbes – will impact the colony's success. Also, honey bees will spend most of their lives within the hive, where horizontal transmission of microbes between beehive environments and the individuals occurs (Anderson et al., 2013). Pesticides, for example, not only cause microbiota dysbiosis on honey bee guts (Kowallik & Mikheyev, 2021; Motta & Moran, 2020), but also accumulate to dangerous levels within hives (Calatayud-Vernich et al., 2018), and may thus affect the different microbial communities present within managed colonies. In addition, food reserves (bee bread, nectar, etc.) can also act as pathogen reservoirs within honey bee colonies and may contribute to pathogen transmission.

Until recently, the majority of bee microbiome investigations focused on the worker bee gut composition and used culture-dependent genome sequencing or 16S rRNA gene amplicon sequencing to characterize the entire bacterial community (Kwong, Medina, et al., 2017). However, shot-gun metagenomic approaches have been proving that the microbiome of honey bees may be much more diverse than previously thought. In a single colony, the microbiome can greatly differ among individuals at the strain-level, and this translates to differences in functional capabilities (Ellegaard & Engel, 2019), which may be also related to the diversity of phages infecting these strains (Bonilla-Rosso et al., 2020; Deboutte et al., 2020). In addition, fungi are also starting to gain more notoriety in microbiome studies, since they are prevalent in bee colonies and may have important interactions with the whole microbial community and the bees (Anderson et al., 2011). Therefore, both amplicon sequencing (16S rRNA for bacteria, ITS for fungi) and shot-gun metagenomics are useful approaches in a microbiome study.

With modified sampling protocols, microbiome studies can be conducted on both tissue samples and hive samples, such as propolis, wax, and honey (Figure 22(B)). Here we present protocols for conducting basic tissue and non-tissue microbiome studies focused on culture-independent approaches that are appropriate for the first screen of bacterial and fungal communities. While some of the methods

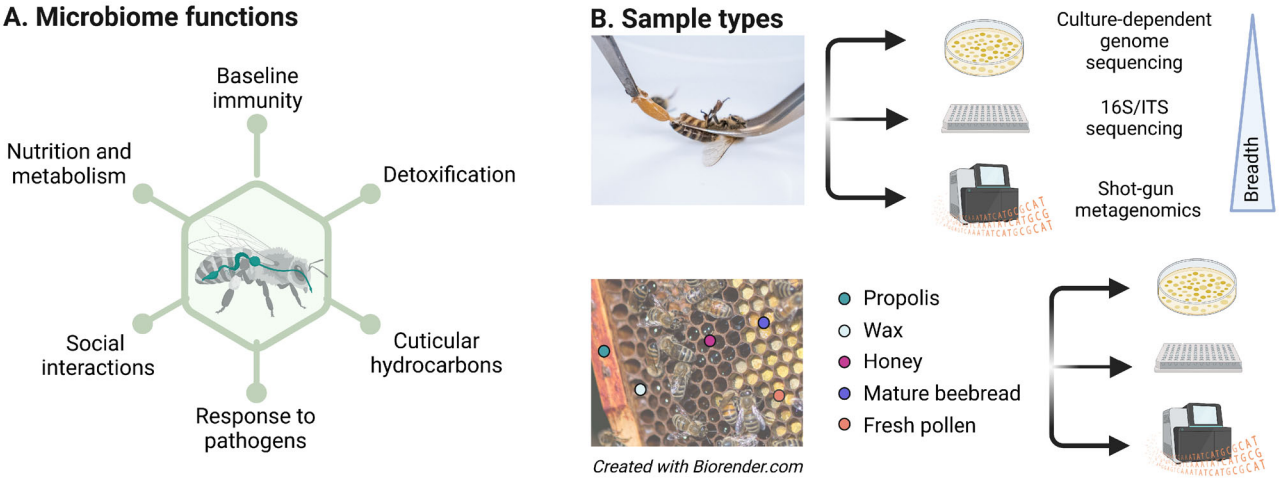

Figure 22. Microbiome overview. (A) Ways the gut microbiome can influence honey bees and colony fitness. (B) Types of samples and sequencing approaches. Results of culture-dependent genome sequencing are limited to taxa amenable to laboratory cultures, while shot-gun metagenomics offers the widest taxa breadth. Note the difference between mature bee bread (shiny surface) and fresh pollen (chalky surface). Photos by Leslie Kennah. Created with Biorender.

described here are equivalent to those described in Engel et al. (2016) and derived from general recommendation for microbiome studies (Hammer et al., 2015), we have included additional sample preparation measures to reduce noise in the data, procedures for sampling hive materials, a worked example of data analysis procedures (including amplicon sequence variant (ASV)-based analysis), links to external resources, and updated commentary on where the field is headed.

10.2. Sampling and DNA extraction

10.2.1. Considerations

10.2.1.1. General

- If you won't be able to use the fresh samples for the DNA extraction or would like to culture microbes from them later, you may follow the procedures for sampling or dissection, then macerate each sample in individual tubes with 1X PBS, add glycerol to a final concentration of 25% and store them at −80 °C until use.
- Change dissection and sample collection instruments (tweezers, scissors, scalpels, spoons, etc.) between samples. If you don't have the option of using individual tools per sample, you can clean instruments between dissections by dipping them in 70% ethanol and flame sterilizing them, or use disposable individual scalpel sterile blades.
- It is a good practice to check for DNA quality and quantity in a 1% agarose gel and Qubit Fluorometer, or directly in an 4200 TapeStation (Agilent Technologies Inc.), following quality parameters stated by the manufacturer.
- Controlling cross-contamination is essential. Spin tubes before opening them, especially after incubation steps, to avoid contamination due to any drops inside the caps.
- You must include negative controls along with the whole experiment: DNA extraction, PCR (in the case of amplicon sequencing), library preparation, and sequencing. These negative controls should be obtained following the exact same protocol as the regular samples, but substituting sample by the buffer. This will allow you to identify contaminants in your sample post-sequencing.
- In addition to relative quantities, you can also recover absolute quantities of microbiome members by using a spike-in standard control containing accurately quantified cell numbers of specific microbe combinations.

10.2.1.2. Tissue sample handling

- If your investigation will focus on the microbes associated with internal organs, you may consider surface-sterilizing the larvae or adult bee to ensure you do not contaminate your sample. However, this is a time-consuming step that may not have a significant impact on results (Hammer et al., 2015).
- The same protocol makes use of commercial kits and suggests specific sequencing platforms for the purpose of presenting a complete procedure. However, while certain steps are essential for reliable microbiome characterization – as indicated in the protocol – the materials used can be replaced according to availability.

10.2.1.3. Hive material sample handling

- Considering the multiple sources of pollen and saliva in bee bread, the pooling of multiple cells ($\geq$10 cells) is recommended to obtain a representative sample. If possible, the targeted cells to be used for DNA extraction should be located on both sides of the comb and on different comb edges.
- Bee bread physicochemical characteristics change over time and affect the bee bread microbiome. Thus, sampling bee bread cells of similar ages is also recommended. Recently packed pollen cells will have separated layers and a porous, chalky texture, crumbling easily. Older bee bread will still have various layers, but with a more waxy and shiny texture.
- For sampling hard surfaces, e.g., the hive entrance, we recommend using cotton-headed swabs, which facilitate in-field sampling, sample transportation, and storage. The drawback of this sampling method is the low DNA load, often paired with low absorbance scores ($<$10 ng/µl).

10.2.2. Protocol for tissue samples

10.2.2.1. Materials

- Kits: DNeasy Blood & Tissue Kit (QIAGEN) and complementary reagents.
- Equipment: Dissection microscope, bead-beating tissue homogenizer (e.g., FastPrep-24™ 5G, (MP Biomedicals); or Precellys-24, (Bertin Technologies)), microtubes, pipettes, and tips (10–1000 ul).
- Reagents: 70% ethanol, 1X PBS, 0.1% sodium hypochlorite solution and 0.22 µm filtered or autoclaved water (optional).
- Disposable dissection tools or materials for quick sterilization of tools between samples (70% ethanol and a flame).
- Sample tubes suitable for a bead-beating homogenizer (e.g., Lysing Matrix E tubes, (**MP Biomedicals**)).

10.2.2.2. Dissection methods

1. Clean and sterilise bench and tools with 70% ethanol.

2. Rinse the bee's body for 3 min in 0.1% sodium hypochlorite and then 3 times in filtered, autoclaved water (optional; see Considerations).
3. Dissect each tissue sample with a new, sterile microdissection tool to avoid cross-contamination. See Carreck et al. (2013) for guidance on dissecting specific tissues.
4. Place the tissue sample in a homogenization tube and proceed to lysis (Section 10.2.2.3.). TIP: If you are only interested in the bacterial fraction, a lysozyme treatment for bacterial cell wall lysis before DNA extraction may also be a good lysis option.

10.2.2.3. DNA extraction methods

5. Add your sample to a Lysing Matrix E tube with 400 µL of 1X PBS. Homogenize in FastPrep-24™ 5G for 45 s at speed 6 m/s.
6. Transfer 180 µl of the sample to a new tube, add 20 µL of proteinase K from DNeasy Blood & Tissue Kit (QIAGEN) and incubate the samples for 2 h at 56 °C. Vortex occasionally during incubation.
7. Add 4 µl RNase A (100 mg/ml), mix by vortexing, and incubate for 2 min at room temperature (~22 °C).
8. Continue by following the recommended steps of the DNeasy Blood & Tissue Kit (QIAGEN) manufacturer's protocol for "Purification of Total DNA from Animal Tissues (Spin-Column Protocol)." For a small amount of initial tissue (e.g., 1 bee gut) elute the DNA in the final step with 2 × 30 µL of AE buffer to improve yields.
9. Store the DNA samples at ≤ ×20 °C until ready for sequencing.

10.2.3. Protocol for sampling hive materials

Bee bread is a sugar-rich mixture of packed pollen, nectar, honey, and enzymes derived from the saliva of honey bees (Anderson et al., 2011). Bee activity and lactic acid bacteria acidify this mixture over time, resulting in decreased bacterial yet increased fungal populations (Disayathanoowat et al., 2020). The bee bread bacteriome is dominated by Proteobacteria and Actinobacteria (Anderson et al., 2011; Disayathanoowat et al., 2020; Muñoz-Colmenero et al., 2020).

The hive entrance, on the other hand, acts as a barrier, separating the hive interior and exterior. Thus, this hive niche can be used to detect and measure organisms entering or leaving beehives. However, there is scarce knowledge regarding the microbial diversity and composition of this beehive niche.

10.2.3.1. Materials

- Kits: QIAamp DNA Mini Kit (QIAGEN)
- Equipment: thermal mixer, vortex, and centrifuge, pipettes, pipette tips, microtubes, chemical hood (bee bread samples only), sterile cotton swabs (hive entrance samples only)
- Reagents: 1× PBS, 96–100% ethanol (to prepare kit solutions)
- For bee bread samples only: phenol:chloroform:isoamyl alcohol (25:24:1, *v/v*), chloroform (>99%)

10.2.3.2. Methods for bee bread sampling and DNA extraction

1. In a comb piece containing bee bread, locate a non-broken and preferably full cell.
2. Depending on sample consistency, introduce either tweezers depth inside the cell (at each side of the bee bread) or a large pipette tip (i.e., cut 1000 µl tips) through the center.
3. Pull out the bee bread carefully, shaking gently, and introduce it into a clean tube large enough to hold pooled samples (2–15 mL). Break the bee bread inside the tube to facilitate subsequent steps
4. Repeat steps 1–4 until an appropriate number of cells are obtained.
5. Add 1 mL of 1× PBS per cell (i.e., 4 mL to a tube containing samples from 4 cells) and homogenize (pipetting) until the sample is uniformly suspended.
6. Transfer 200 µl of sample to a new tube. The remaining sample can be stored at −20 °C (−80 °C, ideally) in case the extraction must be repeated.
7. Add 40 µl of proteinase K and 400 µl of Buffer AL, vortex (~15 seconds) and incubate at 56 °C (mixing at 600 rpm) for 1 hour.
8. Now, work in a fume hood. Transfer the supernatant to a new tube, add 600 µl of phenol:chloroform:isoamyl alcohol, vortex intensely, and centrifuge for 15 min at 14,000 rpm. NOTE: For lab space in which hazardous agents phenol:chloroform can not be used, we recommend the usage of commercial kits dedicated to microbiome DNA extraction as an alternative
9. Transfer the supernatant to a new tube, add 600 µl of chloroform, vortex intensely, and centrifuge for 5 min at 14,000 rpm.
10. Transfer the supernatant to a new tube, add 400 µl of 96–100% ethanol, vortex briefly (~15 seconds), and spin.
11. Working in the fume hood is no longer necessary. Transfer the sample (supernatant from step 10) to the column (no more than 700 µl) and follow the established DNA extraction protocol for the kit, until eluting the DNA in 60 µl of buffer AE.

12. Store the DNA samples at $\leq -20\,^{\circ}C$ until ready for sequencing.

10.2.3.3. Methods for hive entrance sampling and DNA extraction

1. In order to collect microorganisms stuck to the hive entrance or door, use sterile cotton swabs to scrub the surface. To maximize surface covering, the entrance should be swabbed by scrubbing left and right several times (~6) per swab tip (using 3–4 swabs per hive).
2. Cut the heads of two cotton swab samples, deposit in a 2 mL tube, and add 800µl of PBS.
3. Add 20 µl of protease and 400 µl of AL Buffer and vortex.
4. Incubate at 56 °C and 900 rpm for 1.5 hours.
5. Transfer the liquid to a new tube (tube A) and add 20 µl of protease plus 400 µl of AL Buffer to the original tube (tube B).
6. Incubate both tubes A and B at 56 °C and 900 rpm for 1.5 hours.
7. Add 400 µl of 96–100% ethanol and vortex briefly (~15 seconds).
8. Transfer the sample to the column (no more than 700 µl) and follow the established DNA extraction protocol, until eluting the DNA in 100 µl of buffer AE. TIP: Using a larger elution volume increases sample recovery.
9. Store the DNA samples at $\leq -20\,^{\circ}C$ until ready for sequencing.

10.3. Amplicon sequencing

Gene-based markers for bacterial and fungal identification, also known as amplicon sequencing or metabarcoding, is often performed through bacterial- or fungal-specific marker gene sequencing. 16S rRNA gene amplification is the most common for bacteria, while the 18S rRNA gene or the internal transcribed spacer (ITS) are used for fungal analysis (Romero et al., 2019). The 16S rRNA gene is constituted by 9 hypervariable regions (V1–V9) surrounded by conserved sequences (Chakravorty et al., 2007), which are conserved yet variable enough to be differentiated in most bacterial species. The eukaryotic 18S rRNA gene is not as conserved as its prokaryotic equivalent and often results in unidentified taxa (Frau et al., 2019). Longer ITSs (ITS1, ITS2, and the 5.8S rRNA gene), however, have been successfully used to identify fungal gut species in honey bees (Decker et al., 2023; Nguyen & Rehan, 2022; Wen et al., 2017; Yun et al., 2018). The approach described below provides a generic guide for how to conduct amplicon sequencing for microbiome analysis, for which any of the above-mentioned targets could be used. Important differences would be the design of the primer (specific to the target) and the amplification conditions.

10.3.1. Considerations

- The choice of primer can alter diversity estimates and may alter the annealing efficiency for certain templates. Due to the sequence-length variation of ITS (Frau et al., 2019), it is important to carefully choose the pair of primers and sequencing platforms to use. We suggest following the Earth Microbiome protocols to choose primers for the amplification of 16S rRNA, 18S rRNA, and ITS barcode genes. Primers can be ordered directly from Integrated DNA Technologies (IDT) or Eurofins.
- We recommend buying the primers with adapters and barcodes included to reduce the handling and library synthesis cost downstream. For paired-end read sequencing in Illumina platforms, we also recommend targeting regions that will ensure a sequence overlap of at least ~50 bp.
- Conduct PCR for each primer pair in triplicate for each sample. These triplicate samples will be pooled at the end of amplification and serve to reduce random PCR artifacts across your reactions. Use a consistent quantity of DNA input (e.g., 200 ng).
- Use a proofreading high-fidelity polymerase (e.g., Phusion® High-Fidelity PCR Master Mix with HF Buffer, New England Biolabs, Inc.) to limit spurious artifactual diversity introduced in your amplification of the genes.
- Negative control samples should always be sequenced. Some bacterial taxa are known contaminants derived from sample processing (e.g., DNA extraction kits) (Salter et al., 2014), while different sequencing platforms might favor amplification of specific microorganisms.
- Batch effects cause significant differences across experiments, unrelated to biotic factors. These effects are major issues in microbiome data, and common when processing samples using multiple sequencing runs. Thus, it is a good idea to always select a set of samples to use as a reference, which will be added to all sequencing runs of an experiment.
- Amplicon sequencing depth will strongly depend on the expected diversity of the gut microbiome (more diversity requires a higher depth to sample the community). Considering the published work on the bee microbiome, for example, we would recommend a first sequencing depth of 5,000 reads/sample.
- Demonstrated examples for bacterial and fungal amplicon sequencing can be found on the Illumina website.

10.3.2. Materials for amplicon sequencing

- Kits: PureLink PCR Purification Kit (Invitrogen), Quant-iT (Invitrogen), Nextera XT DNA Library Preparation Kit (Illumina), MiSeq Reagent Kit v2 (Illumina)
- Equipment: PCR thermocycler, fluorescent spectrophotometer (compatible with 96-well plates), MiSeq instrument (Illumina) agarose gel electrophoresis apparatus and associated reagents
- Reagents: Phusion® High-Fidelity PCR Master Mix with HF Buffer (New England Biolabs, Inc). Forward and Reverse primers to target the particular gene of interest (e.g., 16S rRNA, 18S rRNA, ITS)
- PCR plates

10.3.3. Methods for amplicon sequencing

1. Mix a defined quantity of DNA with 1× of Phusion® High-Fidelity PCR Master Mix with HF Buffer (New England Biolabs, Inc), forward and reverse primers to a final concentration of 0.5 μM, in a final reaction volume of 50 μL.
2. In the thermocycler run an initial denaturation at 98 °C for 2 min, followed by 30 cycles of 98 °C denaturation for 10s, 55 °C annealing for 30 s and 72 °C extension for 30 s, with a final extension step at 72 °C for 10 min. TIP: Be attentive to primer melting point requirements and change the annealing temperature if needed.
3. Check for the amplicon size in a 1% agarose gel and then purify the PCR product with PureLink PCR Purification Kit following the manufacturer's recommendations. If you did not include barcodes in your primer synthesis, you must prepare libraries with the Nextera XT DNA Library Preparation Kit, following the manufacturer's instructions, using one barcode per sample.
4. The amplicons produced as part of this protocol must be normalized by concentration and pooled for sequencing. If you will be pooling your samples yourself, we recommend using the Quant-iT kit, as per manufacturer instructions.
5. After pooling your sample such that the same amount of DNA from each amplicon is added to your sequencing run, you will clean up your PCR amplicons using the PureLink PCR Purification Kit again.
6. These amplicons should be sequenced using 2 × 250 paired-end reads on a Miseq platform using the MiSeq Reagent Kit v2.

10.4. Microbiome data analysis

Regardless of whether 16S, 18S, or IST amplicons are used, the general principle is that microbial taxonomy can be inferred by comparing the gene fragment sequence to a specific database. Concurrently, relative abundances can be calculated based on the number of amplicon copies. Below we provide guidance on strategies and softwares that can be used to complete this type of analysis.

10.4.1. Recommended software

There are currently several pieces of software used to process amplicon data, either by inferring operational taxonomic units (OTUs) or amplicon sequence variants (ASVs). While OTUs represent a cluster of multiple and similar sequences, for ASVs the differences in nucleotides will result in an unique variant and a more detailed picture of the diversity. We recommend here the Mothur pipeline (Schloss, 2020) for quality control of your reads, contig formation, chimera removal, alignment to the 16S rRNA gene, and generation of OTUs and ASVs. For taxonomy, Mothur uses Bayesian analysis of kmer profiles, returning the most likely match with the database. All these steps are detailed in the standard operating procedure (SOP) (Schloss, 2020).

Commonly used databases are SILVA and UNITE, for bacteria and fungi respectively. Several other more curated databases are also available for the study of bee species, such as RDP for annotated bacterial and archaeal 16S rRNA sequences and fungal 28S rRNA sequences (Cole et al., 2014), and BEExact for bacterial 16S rRNA sequences often found in bee species (Daisley & Reid, 2021). The choice of database is important, as it can lead to errors in taxonomic placement if sequence representatives from your environment are misidentified or absent (Newton & Roeselers, 2012).

The next steps may include statistical analyses and plotting results, for which you can continue to follow the Mothur SOP or the QIIME 2 pipeline (see Section 10.4.2). In general, we recommend measuring alpha diversity to generate rarefaction curves describing the number of OTUs or ASVs as a function of sampling effort, and beta diversity to compare samples' community structure. Distance matrices can be visualized using the Principal Coordinates Analysis (PCoA) or the Non-metric multidimensional scaling (NMDS) plots. You can also test for the microbiota dynamics with a Permutational Multivariate Analysis of Variance (PERMANOVA) using factors of your choice (e.g., caste, apiary, season, treatments). (Engel et al., 2013) has more information on these exploratory techniques. The Mothur wiki (https://mothur.org/wiki/) has extensive information about data processing in their manual. Here, we provide an alternative worked example using QIIME 2.

10.4.2. Guidance on the data analysis methods: an example with QIIME 2

QIIME 2 (Bolyen et al., 2019) and its community (**QIIME 2** forum; https://forum.qiime2.org/) offer standardized pipelines for the analysis of microbial communities. **QIIME 2** core concepts, hardware/software/metadata requirements, installation, (re)activation, and core applications are easily accessible in Qiime2 documentation (**QIIME 2** docs), and not explained in this protocol. Keeping track of the official **QIIME 2** docs is recommended since it will reflect the latest **QIIME 2** release. Herein, we will only present the ITS-specific steps as an example. We will follow the **Casava** 1.8 protocol for paired-end demultiplexed sequences.

10.4.2.1. Importing data. Data importation in **QIIME 2** is dependent on the data format. FASTQ documents containing single-end or paired-end sequences are the most common raw data. Forward and reverse sequences are usually referred to as R1 and R2, respectively, while barcode sequences are named I1 (index). FASTQ documents can be either multiplexed (one file per sequence "type," I1, R1, and/or R2) or demultiplexed (one file per sample, with its corresponding R1 and/or R2 sequences). The most common demultiplexed format is **Casava**, obtained through Illumina's Casava software. Although **QIIME 2** has implemented commands for the easy importation of the most common data formats, any other type of data format (supported by **QIIME 2**) will have to be imported through "Fastq manifest" (https://docs.qiime2.org/2021.8/tutorials/importing/?highlight=casava) or similar pipelines.

Step 1. To import the data, enter the following commands into the terminal:

```
qiime tools import\
- -type "SampleData[PairedEndSequencesWithQuality]"\
- -input-path casava-18-paired-end-demultiplexed\
- -input-format CasavaOneEightSingleLanePerSampleDirFmt\
- -output-path demux-paired-end.qza
```

10.4.2.2. Non-biological sequence removal. Imported demultiplexed sequences contain "non-biological" sequences (i.e., primers), which have to be removed. Hypervariable-length amplicons such as ITS can contain 4 non-biological sequences: F primer at the beginning of F sequences, reverse complementary sequence of the R primer at the end of F sequences, R primer at the beginning of R sequences, and reverse complementary sequence of the F primer at the end of R sequences.

Step 2. To remove these non-biological sequences, execute the following commands:

```
qiime cutadapt trim-paired\
- -i-demultiplexed-sequences demux-paired-end.qza\
- -p-adapter-f GCATATCAATAAGCGGAGGA\#Reverse complementary
  sequence of Rp.
- -p-front-f GTGARTCATCGAATCTTTG\#Forward primer (Fp) sequence:
  fITS7
- -p-adapter-r CAAAGATTCGATGAYTCAC\#Reverse complementary
  sequence of Fp.
- -p-front-r TCCTCCGCTTATTGATATGC\#Reverse primer (Rp) sequence:
  ITS4
- -p-error-rate 0.1\#Allowed error-rate (default)
- -o-trimmed-sequences trimmed.qza\
- -verbose
```

10.4.2.3. Sequence quality control (denoising) Denoising in **QIIME 2** can be performed through **DADA2** (Callahan et al., 2016) or **Deblur** (Amir et al., 2017). **DADA2** produces amplicon sequence variants (ASVs) while Deblur gives sub-operational-taxonomic-units (sub-OTUs or sOTUs). Herein, we will follow the **DADA2** protocol, wherein paired-end reads are joined and denoised simultaneously. Two parameters are needed: trimming positions (starts of F and R reads) and truncating positions (ends of F and R reads). Both parameters can be determined by visualizing the demultiplexed data and checking the interactive quality plots, which contain quality score values per sequence base.

Step 3. Determine trimming and truncating positions:

```
qiime demux summarize\
- -i-data trimmed.qza\
- -o-visualization trimmed.qzv
```

Optimal parameters result in merged reads of good quality. F and R have to be long enough to merge, and merged sequences have to be good enough to pass the filtering threshold of DADA2.

Step 4. Trim primer sequences:

```
qiime dada2 denoise-paired\#Perform Dada2 denoising
- -i-demultiplexed-seqs demux-paired-end_trimmed_def.qza\
- -p-trim-left-f 13\#Trim Forward sequences in 13rd position
- -p-trim-left-r 0\#Do not trim Reverse sequences
- -p-trunc-len-f 0\#Do not truncate the end of Forward sequences
- -p-trunc-len-r 220\#Truncate Reverse sequences in 220th position
- -o-table dada2/table.qza\
- -o-representative-sequences rep-seqs.qza\
- -o-denoising-stats denoising-stats.qza
```

Step 5. Visualize denoising results:

```
qiime metadata tabulate\
- -m-input-file denoising-stats.qza\
-o-visualization denoising-stats.qzv
```

From this point on, paired-end and single-end data are analyzed following the same steps, for all microbial communities. In order to assign taxonomy

ID to fungal sequences, the UNITE database is used, which requires to be first trained. It is recommended to train the UNITE classifier as follows:

- Use non-extracted full ITS sequences from the developer UNITE database (**QIIME**-compatible release).
- Use the q2-itsxpress plugin, which permits extraction of ITS domains from input data (sequences). Then, extracted sequences can be compared to the standard UNITE database.

Available tutorials for diversity and compositional analysis can be followed at the **QIIME** docs website (https://docs.qiime2.org/2022.11/) and a detailed pipeline can be found at (Estaki et al., 2020).

10.4.2.4. Removing biological contamination. Despite best efforts, biological contamination of samples with microbes not belonging to the sample is still possible. Excluding suspected contaminants from downstream analyses is encouraged. For example, excluding suspected contaminants from downstream analyses is appropriate if, according to the sequencing of blank controls, contamination by laboratory reagents or the laboratory environment is suspected. If necessary, there are multiple softwares for removal of contaminant sequences, such as the **R** package **decontam** (Davis et al., 2018; Salter et al., 2014).

10.5. Applications and limitations

In a honey bee colony, each individual plays a specific yet adaptive role, and changes in the microbiome composition (dysbiosis) of even just some individuals may influence the colony's success or failure. Thus, the microbiome should be evaluated more often in bee health studies, as it is known to be affected by pesticides (Kakumanu et al., 2016), pathogens (Paris et al., 2020), food restrictions (Castelli et al., 2020), and even change in environmental cues (Hammer et al., 2021). The majority of the studies have focused efforts on characterizing worker bee gut microbiota, but it is important to incorporate investigations of other castes, tissues, developmental stages, and colony environments.

The honey bee colony is ultimately a superorganism (Moritz & Southwick, 2012), and the microbes harbored in different parts of the colony can also play a direct role in colony fitness or may serve as microbe reservoirs. The microbial community does not exist in isolation, and we encourage readers to consider not only the simple characterization of the communities, but their interactions with social evolution (Liberti et al., 2022; Vernier et al., 2020), development (Hammer & Moran, 2019), diapause (Mushegian & Tougeron, 2019; Santos et al., 2019), and the brain-microbiome axis (Zhang, Mu, Cao, et al., 2022; Zhang, Mu, Shi, et al., 2022).

Amplicon sequencing approaches do, of course, have limitations, one of which is that analyzing hyper-variable regions is not equal sensitive for all bacteria (one hypervariable region is insufficient to differentiate all bacterial species). Most OTUs or ASVs will thus not have the sequence resolution to taxonomically identify the microbes to genus or species, much less strain (Callahan et al., 2017). Even so, this approach is still useful for microbial community characterization and for observing major changes in it.

Different strains of microbes may fill different functional niches, and investigating patterns of strain variation is a growing area of interest. Primers for non-marker genes have already been developed to detect diversity at the bacterial strain level, including the genes minD (Powell et al., 2016), guaA, and gluS (Bobay et al., 2020) for Snodgrassella alvi strain composition, and pflA and rimM for Gilliamella spp. strain composition. This method was termed metagenomic amplicon strain typing (MAST). Although an interesting approach focused on specific members of the microbiome, there are no similar published studies for fungi. Otherwise, to describe the strain-level diversity of the entire microbial community, a shotgun metagenomics approach would be more appropriate (Ellegaard & Engel, 2019), but it is more expensive and time-consuming for analysis. However, those who utilize this method are rewarded with not only strain-level diversity, but also host genotypes and gene sequences of the microbes, potentially allowing for functional insights.

Conversely, using the amplicon sequencing approach, there is limited knowledge to gain regarding the roles of microbiome members, since there is no information regarding their gene repertoire and metabolic activity. Of course, functions can be predicted based on general knowledge of the described species or sequenced genome using software as PICRUSt (Langille et al., 2013), Tax4Fun2 (Wemheuer et al., 2020), and FUNGuild (Nguyen et al., 2016), but strains may differ with regards to their functional abilities; therefore, these predictions are tenuous at best. One way to alleviate this concern would be to combine culture-dependent approaches to conduct *in vitro* and *in vivo* experiments – confirming a microbes' role and interactions. This kind of functional validation of predictions remains one of the field's biggest challenges.

The field of honey bee microbiome analyses is still underway. In particular, the non-bacterial communities within honey bee guts are not well characterized, with a lack of consensus regarding fungal

communities (Hroncova et al., 2015; Khan et al., 2020) and ubiquitous viral communities (Bonilla-Rosso et al., 2020; Kadlečková et al., 2022). And, despite the characterization of the bacterial profiles of honey bee guts, much is still unknown regarding its impact in honey bee colony fitness, such as how microbial communities influence host physiology or how spatial specialization of microbes within colonies occurs (Copeland et al., 2022; Powell et al., 2021; Zheng et al., 2017).

11. Data management and open access sharing

Past the excitement from the access to large amounts of sequencing data for numerous honey bee species, populations, individuals, and tissues, researchers may encounter challenges in determining how datasets were obtained. This problem is not inherent to the honey bee community and inconsistencies appear often in many organisms (Gonçalves & Musen, 2019). In the context of honey bee omics research, we have identified two main challenges that hinder comparative studies: (1) the lack of details and access to bioinformatics scripts used and (2) the heterogeneity of metadata associated with open access data. These problems are avoidable with improved practices and standards. Addressing these challenges through standardization and transparency will promote reproducibility and facilitate data comparability and integration.

There are several global and honey bee-specific databases that exist and provide platforms for data submission and search. These databases serve as valuable resources for researchers to access and contribute to the wealth of information available on honey bees. Across the tree of life, the core databases NCBI (National Center for Biotechnology Information), DDBJ (DNA Data Bank of Japan), and ENA (European Nucleotide Archive) are widely recognized as major repositories for biological data. Such databases are now fed by ambitious initiatives that aim to sequence and analyze the genomes of a vast number of species, including those beyond model organisms such as the Earth Biogenome Project, Darwin Tree of life, and i5K. These initiatives contribute to the growth of genetic data resources and can indirectly benefit honey bee research.

In addition to the global databases mentioned earlier, several functional genetic databases play important roles in specific model organisms and functional genomics research. While they may not directly contain honey bee datasets, they can still be utilized as proxies for comparative analysis. Some relevant databases include: Flybase, Beetlebase, BUSCO/OrthoDB (Benchmarking Universal Single-Copy Orthologs) and Gene Ontology (GO).

Apis-specific databases also exist and have been developed to specifically focus on honey bee microbiome research and provide curated sets of honey bee microbes and associated tools with Bee-exact and Holobee. Another example is the beenome100 project database which serves as a valuable genomic resource and establishes a comprehensive phylogenomic framework for *Apis* genus by aiming at generating reference genomes for 100 U.S. bee pollinators species.

11.1. Metadata standardization

While peer-reviewed journals require that any genetic data are shared in open-access, no rigorous quality control of the metadata submitted is ensured. Regardless of the reader's expertise level in submitting genomic data, we recommend following the best practices listed by the journal Scientific Data ("Promoting best practice in nucleotide sequence data sharing," 2020) in submitting metadata to NCBI, DDBJ, ENA, and other databases. This section does not intend to replace the multiple and comprehensive resources available guiding data and metadata submission, and we urge to carefully adhering to these standards (i.e., for NCBI BioProject).

11.1.1. Common problems with Apis-related BioProjects

Often, a single study can be associated with a BioProject (i.e., BioProject 1 = X individuals SNPs data, BioProject 2 = Y Transcriptomes) or eventually an umbrella project gathering several related projects (i.e., BioProject UMBRELLA [BioProject 3 = Proteomics data for 100 individuals + and BioProject 2 = Gene expression data for 50 individuals]). One common mistake is the absence of a clear, unique title and description.

Among the 517 BioProjects strictly related to *Apis* honey bee available on NCBI, we found that most did not have associated publications or proper descriptions (Search: "*Apis mellifera*"[Organism] and manual sorting). This can make it difficult for the reader to know (1) how the data were generated and (2) which institute or team could be contacted to reach out on the origin of the data. Thus, we suggest that researchers follow the subsequent best practices for BioProject submission:

Best practices for BioProjects:

- Title should be as explicit and descriptive as intended for publication.
- Description should include details regarding (1) the project aim, (2) a general overview of the *Apis* species, subspecies, populations, individuals, and sex targeted, (3) the material origin (e.g., tissue specific vs. whole body, single individual vs. pool), and (4) which sequencing platform and library approach were used.
- In case the data were released prior to publication, we recommend associating a link to the lab in charge with the institute/unit/team indicated as submitter.
- Indicate a data usage statement in case of early release to inform users about possible embargo.
- Contact database curator to give the DOI of the associated published report or paper.

11.1.2. Common problems with Apis*-related BioSamples*

From our survey, most of the inconsistency and variability observed in metadata submitted for *Apis* honey bee projects was observed in BioSample (which contains important metadata regarding the source of the sample). A BioSample is essential for cross-scale comparative studies and data mining. Several studies in *A. mellifera* population genetics (see Section 4.6) have generated new genome-wide data but also compared with former studies from a different source in their analysis (Cridland et al., 2017; Dogantzis et al., 2021; Shi et al., 2020; Tihelka et al., 2020). Such integrative studies are predicted to increase in the future as they give a broader and global perspective on honey bee evolution and biology.

We found that 16,855 BioSamples related to honey bees and *Apis*-associated environmental organisms (except Acari mites and hive insect pests) were registered in public databases. Among them only 60% had attribute fields about the specimen's sex, and 53% had geographic location details. Aside from incomplete data, in our survey we encountered problems related to variable orthograph even with fields as straight-forward as "sex" attribute (Table 1). While the absence of such data can be understandable and common with historical or third-party sampling, we urge future submissions to include at least the following details for *Apis* standard research.

Best practices for BioSamples:
For *Apis* specimen data, use the invertebrate submission package (https://www.ncbi.nlm.nih.gov/biosample/docs/packages/Invertebrate.1.0/) which requires the collection date, the geographical origin, and the tissue used for sequencing as mandatory attributes.

For *Apis* environmental and metagenomic-transcriptomic data, use the MIMS: metagenome/environmental, host-associated; version 5.0 package (https://www.ncbi.nlm.nih.gov/biosample/docs/packages/MIMS.me.host-associated.5.0/).

In the specific attributes field, we recommend researchers to include the following additional information, although it is not mandatory according to the NCBI submission system.

Attribute	Recommendations
"Breed" or create an additional column "Subspecies." Avoid using "isolate" for this purpose and instead use this attribute for the sample name.	When known, include the subspecies or strain level of the honey bee studied (e.g., *Apis mellifera* unicolor, *Apis cerana* japonica, *Apis mellifera* "Buckfast")
"biomaterial_provider" and "collected_by"	Acknowledge the person in charge of the sampling with Name + Affiliation
"dev_stage"	Indicate if the sample was related to an egg, larvae, pupae or adult bee
"lat_lon" and "geo_loc_name"	Give as many details on the sampling location using geo-coordinates in decimal format, which could be use for spatial studies or to guide future sampling. In some case, the exact geocoordinates can be lost but we recommend to use at least approximate ones (e.g., centroid of a city or locality).
"treatment"	For transcriptomics and epigenomics studies, the various treatment conditions should be included in the metadata.
"host" and "tissue"	For metagenomics, metatranscriptomics and microbiomic, include the host species and the origin of the tissue sampled

Ideally, supplemental data table should be provided with the sample name, library ID, and all the aforementioned attributes. This would allow future users to cross-check the information with the analysis results, which may not be explicitly stated in the raw data deposited to the database. Examples of such information include population structure assignment, patriline, omic profile, and other relevant details.

The retroactive correction and update of both BioProject and BioSamples content is often easy and requires the original submitter to contact a curator from the database of initial submission For instance, if you submit your Sequence Read Archive (SRA) data and metadata to DDBJ, you can provide a list of all corrections for each accession to a DDBJ curator. These changes will then be automatically transferred to other databases, such as NCBI, within a

Table 1. Survey results of BioSample attributes.

Orthographs for "sex" attribute	Number of SRA
Female	3629
femle [sic]	1
Male	1966
MISSING	79
None	6
Not applicable	6
Not collected	16
Not determined	6
Pooled male and female	418
Sterile female	12
Worker	22
(No entry)	7629

short period of time. Since these databases communicate on a daily basis, the changes can be implemented swiftly.

11.2. Sharing pipelines and scripts

The burst and diversification in the development of software, **R** packages, pipeline and servers to analyze omics data can quickly lead users to go down a bioinformatic rabbit hole to make the "best" choice. Additionally, the frequent emergence of comparative studies promoting new tools adds to the difficulty of decision-making. In this chapter, we have proposed up-to-date and efficient standard pipelines to guide users in their methodological workflows. However, we acknowledge that these pipelines may require future revisions, similar to how the present chapter updates information from Evans et al. (2013). We aim to provide a valuable resource that evolves alongside advancements in the field.

As progress is made and new analytical workflows are adopted, we invite the honey bee research community to contribute to the data processing standardization by sharing their bioinformatics methods and custom scripts. While sharing sequencing data is mandatory in peer-reviewed journals, the same level of detail is often not required for downstream analysis. Minimal descriptions are often given in the "Material and Methods" sections regarding parameters for each software or script function. Depending on the journal's requirements, it may be necessary to share analysis output data (e.g., VCF files, LFMM, FASTA) and scripts through external repositories associated with a DOI.

Noteworthy initiatives, such as the **SeqApiPop** population genomics study, have inspired changes by publishing detailed codes on collaborative platforms like GitHub (github.com/avignal5/SeqApiPop) alongside their results (Wragg et al., 2021). Similar initiatives using interactive **R** and shell markdowns have been carried out for honey bee microbiome (Kowallik & Mikheyev, 2021; Liberti et al., 2022) and transcriptomics studies (Holman et al., 2019; Warner et al., 2019). In addition to enabling reproducibility and standardization, we advocate that open access to the code scripts (default and personalized) serves as a valuable teaching tool, particularly tailored to honey bees. To facilitate the growth of such resources, we recommend the following best practices:

Best practices for sharing pipelines and scripts:

- Whole genome sequencing and population genomics studies should strive to make their data freely available through international open-access repository such as Dryad or Zenodo.
  Content: VCF file, Demographic analysis, FASTA alignments
- Custom scripts (bash, Python, awk) and pipelines created with Workflow Management Systems (such as Snakemake, Nextflow, Galaxy) should be made freely accessible to a cloud-based platform with version control, such as GitHub (or alternatives: GitLab, Bitbucket, SourceForge, etc.) or Wiki.
  Content: Markdown, codes, notes, input files (small sizes)
- State clearly in publication the source and link to data and code availabilities

While it is challenging to predict which platforms will remain relevant over the next decade, it should be noted that free access to well-curated scripts can lead to rapid dissemination across various formats and applications if we promote a "culture of reproducibility" (Peng, 2011). Despite natural skepticism about the longevity of experimental software and pipelines, the trend of open-source and community-driven of these tools suggests that many will remain valuable resources (Ewels et al., 2020).

When selecting a community-curated pipeline, users should consider the following: (1) How often has this pipeline been cited and used? (2) Does the closed/open issues ratio or forum activity indicate strong developer technical support? (3) How frequently is the pipeline updated (consider the dates and number of releases)? (4) Is there an active user community providing support? Additionally, with the rise of package manager such as Conda or containerization technologies like Docker and Singularity, it's important to consider if the pipeline is available in a containerized format, as this can significantly improve the reproducibility and ease of its usage across different computing environments.

12. The future of *Apis* omics: biological integration

In the coming decades, we anticipate the emergence of new methods driven by technological advances that will further reduce costs and expand omics applications beyond model species. At the moment, cutting-edge approaches, such as single-cell sequencing and atlas generation (Luecken & Theis, 2019; Misra et al., 2022), spatialomics (Moffitt et al., 2022), and the use of machine learning for various omics analyses (Arjmand et al., 2022; Li, Li, et al., 2022) have been developed and applied for clinical and forensic purposes. These advancements hold great promise for honey bee research, offering unprecedented insights into the evolution of phenotypic plasticity, behavioral profiling within a superorganism, and the origins of eusociality.

However, immediate progress in our understanding of *Apis* honey bee biology, diversity, and evolution can be achieved by the layering of multi-omics data and interpretation (Toth & Zayed, 2021). Genomics, epigenomics, and transcriptomics have seen remarkable growth in honey bee research and provided valuable insights. By integrating multi-

omics data from the same honey bee sample (single cell, tissue, individual or colony), we can uncover the mechanistic and immediate causes behind behavioral changes, such as labor division, as well as responses to various stress factors like pathogen infections or chemical exposure. The combination of multiple omics approaches has been proposed as a toolkit to better characterize and improve honey bee health (Grozinger & Zayed, 2020). Initiatives like the Canadian BeeCSI project are actively working on integrating multiple omics approaches to better understand and promote honey bee health.

The integration of behavioral assays, chemical profiling and metagenomics has shed light on the significance of honey bee host-microbiome interactions in nestmate recognition (Vernier et al., 2020). Leveraging functional genomics and transcriptomics, based on knowledge gained from studying the honey bee microbiome could help engineer innovative pathogen control methods (Leonard et al., 2020). As a last example, new insights in *A. mellifera* social immunity via the discovery of transmissible RNA in shared royal jelly resource, was made possible by the combination of proteomics, transcriptomics and functional genomics (Maori, Garbian, et al., 2019; Maori, Navarro, et al., 2019). Encouraged by these achievements, further multi-omics surveys are anticipated to expand our knowledge to other *Apis* species.

Acknowledgements

The COLOSS (Prevention of honey bee (COlony LOSSes) Association aims to explain and prevent massive honey bee colony losses. It was originally funded through the COST (European Cooperation in Science and Technology) Action FA0803. The COLOSS Association is now supported by the Ricola Foundation – Nature & Culture.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

M.A.T.'s postdoctoral research was supported by a fellowship from the Japan Society for Promotion of Science (JSPS) [P19723] and by the BPRI funded by the NSF-BII award [2021795]. Am was supported by a L'Oréal-UNESCO For Women in Science Fellowship and a Project *Apis m.* grant. Financial support was granted to H.L.-B. by the USDA NIFA Evans-Allen funds [NI171445XXXXG004, NI201445XXXXG018-0001], a NSF-RIA award [1900793], a USDA NIFA AFRI grant [2020-67014-31557], and USDA SARE Grants [ONC19-062 and LNC21-459]. P.C.'s research was funded by the United States Department of Agriculture—Agriculture and Food Research Initiative (USDA-AFRI) and a grant from Project *Apis m*. D.H. was supported by the project MEDIBEES—Monitoring the Mediterranean honey bee subspecies and their resilience to climate change for the improvement of sustainable agro-ecosystems. MEDIBEES is part of the PRIMA program supported by the European Union. D.H. and M.A.P were supported by national funds from Fundação para a Ciência e a Tecnologia (FCT)/MCTES (PIDDAC) to CIMO [UIDB/00690/2020 and UIDP/00690/2020] and SusTEC [LA/P/0007/2021]. J.G. is supported by a predoctoral grant from the Department of Agriculture, Fisheries and Food of the Basque Government. J.G.'s, M.P.'s and I.Z.'s work was supported by the Government of the Basque Country (Research Group IT1571-22). I.L.G.N. and L.C.'s work was supported by an NSF DBI grant [2022049] and NSF IOS grant [2005306].

ORCID

June Gorrochategui-Ortega http://orcid.org/0000-0001-7080-8337
Iratxe Zarraonaindia http://orcid.org/0000-0002-0615-0187

Data availability statement

Datasets associated with the tutorials described in this paper are available at github.com/MaevaTecher/standard-apis-omics. This paper is also available as a wiki online at https://maevatecher.github.io/standard-methods-apis-omics/; doi.org/10.5281/zenodo.14697986.

References

Adema, C. M. (2021). Sticky problems: Extraction of nucleic acids from molluscs. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *376*(1825), 20200162. https://doi.org/10.1098/rstb.2020.0162

Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, *537*(7620), 347–355. https://doi.org/10.1038/nature19949

Aguirre-Liguori, J. A., Luna-Sánchez, J. A., Gasca-Pineda, J., & Eguiarte, L. E. (2020). Evaluation of the minimum sampling design for population genomic and microsatellite studies: An analysis based on wild maize. *Frontiers in Genetics*, *11*, 870. https://doi.org/10.3389/fgene.2020.00870

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., & Mason, C. E. (2012). methylKit: A comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biology*, *13*(10), R87. https://doi.org/10.1186/gb-2012-13-10-r87

Albert, C. H., Yoccoz, N. G., Edwards, T. C., Jr., Graham, C. H., Zimmermann, N. E., & Thuiller, W. (2010). Sampling in ecology and evolution – bridging the gap between theory and practice. *Ecography*, *33*(6), 1028–1037. https://doi.org/10.1111/j.1600-0587.2010.06421.x

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Alghamdi, A. A., & Alattal, Y. Z. (2024). Alterations in histone methylation states increased profusion of lethal(2)-essential-for-life-like (l(2)elf), trithorax and polycomb genes in *Apis mellifera* under heat stress. *Insects*, *15*(1), 33. https://doi.org/10.3390/insects15010033

Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J., C Ewald, J., Fraser, P. D., Giavalisco, P., Hall, R. D., Heinemann, M., Link, H., Luo, J., Neumann, S., Nielsen, J., Perez de Souza, L., Saito, K., Sauer, U., Schroeder, F. C., … Fernie, A. R. (2021). Mass spectrometry-based metabolomics: A guide for annotation, quantification and best reporting practices. *Nature Methods*, *18*(7), 747–756. https://doi.org/10.1038/s41592-021-01197-1

Ament, S. A., Blatti, C. A., Alaux, C., Wheeler, M. M., Toth, A. L., Le Conte, Y., Hunt, G. J., Guzmán-Novoa, E., Degrandi-Hoffman, G., Uribe-Rubio, J. L., Amdam, G. V., Page, R. E., Jr, Rodriguez-Zas, S. L., Robinson, G. E., & Sinha, S. (2012). New meta-analysis tools reveal common transcriptional regulatory basis for multiple determinants of behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(26), E1801–10. https://doi.org/10.1073/pnas.1205283109

Ament, S. A., Wang, Y., Chen, C.-C., Blatti, C. A., Hong, F., Liang, Z. S., Negre, N., White, K. P., Rodriguez-Zas, S. L., Mizzen, C. A., Sinha, S., Zhong, S., & Robinson, G. E. (2012). The transcription factor ultraspiracle influences honey bee social behavior and behavior-related gene expression. *PLOS Genetics*, *8*(3), e1002596. https://doi.org/10.1371/journal.pgen.1002596

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., & Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, *2*(2), e00191-16. https://doi.org/10.1128/mSystems.00191-16

Anderson, K. E., Sheehan, T. H., Mott, B. M., Maes, P., Snyder, L., Schwan, M. R., Walton, A., Jones, B. M., & Corby-Harris, V. (2013). Microbial ecology of the hive and pollination landscape: Bacterial associates from floral nectar, the alimentary tract and stored food of honey bees (*Apis mellifera*). *PLOS One*, *8*(12), e83125. https://doi.org/10.1371/journal.pone.0083125

Anderson, K. E., Sheehan, T. H., Eckholm, B. J., Mott, B. M., & DeGrandi-Hoffman, G. (2011). An emerging paradigm of colony health: Microbial balance of the honey bee and hive (*Apis mellifera*). *Insectes Sociaux*, *58*(4), 431–444. https://doi.org/10.1007/s00040-011-0194-6

Anderson, E. C. (2010). Assessing the power of informative subsets of loci for population assignment: Standard methods are upwardly biased. *Molecular Ecology Resources*, *10*(4), 701–710. https://doi.org/10.1111/j.1755-0998.2010.02846.x

Arad, M., Ku, K., Frey, C., Hare, R., McAfee, A., Ghafourifar, G., & Foster, L. J. (2025). What proteomics has taught us about honey bee (*Apis mellifera*) health and disease. *Proteomics*, *25*(1–2), e2400075. https://doi.org/10.1002/pmic.202400075

Arathi, H. S., Bjostad, L., & Bernklau, E. (2018). Metabolomic analysis of pollen from honey bee hives and from canola flowers. *Metabolomics*, *14*(6), 86. https://doi.org/10.1007/s11306-018-1381-5

Ardalani, H., Vidkjær, N. H., Kryger, P., Fiehn, O., & Fomsgaard, I. S. (2021). Metabolomics unveils the influence of dietary phytochemicals on residual pesticide concentrations in honey bees. *Environment International*, *152*, 106503. https://doi.org/10.1016/j.envint.2021.106503

Arjmand, B., Hamidpour, S. K., Tayanloo-Beik, A., Goodarzi, P., Aghayan, H. R., Adibi, H., & Larijani, B. (2022). Machine learning: A new prospect in multi-omics data analysis of cancer. *Frontiers in Genetics*, *13*, 824451. https://doi.org/10.3389/fgene.2022.824451

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., Zhang, G., & 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

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Ashby, R., Forêt, S., Searle, I., & Maleszka, R. (2016). MicroRNAs in honey bee caste determination. *Scientific Reports*, *6*(1), 18794. https://doi.org/10.1038/srep18794

Avalos, A., Fang, M., Pan, H., Ramirez Lluch, A., Lipka, A. E., Zhao, S. D., Giray, T., Robinson, G. E., Zhang, G., & Hudson, M. E. (2020). Genomic regions influencing aggressive behavior in honey bees are defined by colony allele frequencies. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(29), 17135–17141. https://doi.org/10.1073/pnas.1922927117

Avalos, A., Pan, H., Li, C., Acevedo-Gonzalez, J. P., Rendon, G., Fields, C. J., Brown, P. J., Giray, T., Robinson, G. E., Hudson, M. E., & Zhang, G. (2017). A soft selective sweep during rapid evolution of gentle behaviour in an Africanized honeybee. *Nature Communications*, *8*(1), 1550. https://doi.org/10.1038/s41467-017-01800-0

Baky, M. H., Abouelela, M. B., Wang, K., & Farag, M. A. (2023). Bee pollen and bread as a super-food: A comparative review of their metabolome composition and quality assessment in the context of best recovery conditions. *Molecules*, *28*(2), 715. https://doi.org/10.3390/molecules28020715

Barkovits, K., Pacharra, S., Pfeiffer, K., Steinbach, S., Eisenacher, M., Marcus, K., & Uszkoreit, J. (2020). Reproducibility, specificity and accuracy of relative quantification using spectral library-based data-independent acquisition. *Molecular & Cellular Proteomics*, *19*(1), 181–197. https://doi.org/10.1074/mcp.RA119.001714

Barron, A. B., Maleszka, R., Vander Meer, R. K., & Robinson, G. E. (2007). Octopamine modulates honey bee dance behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(5), 1703–1707. https://doi.org/10.1073/pnas.0610506104

Bartel, J., Krumsiek, J., & Theis, F. J. (2013). Statistical methods for the analysis of high-throughput metabolomics data. *Computational and Structural Biotechnology Journal*, *4*(5), e201301009. https://doi.org/10.5936/csbj.201301009

Bataglia, L., Simões, Z. L. P., & Nunes, F. M. F. (2021). Active genic machinery for epigenetic RNA modifications in bees. *Insect Molecular Biology*, *30*(6), 566–579. https://doi.org/10.1111/imb.12726

Batth, T. S., Francavilla, C., & Olsen, J. V. (2014). Off-line high-pH reversed-phase fractionation for in-depth phosphoproteomics. *Journal of Proteome Research*, *13*(12), 6176–6186. https://doi.org/10.1021/pr500893m

Behura, S. K., & Whitfield, C. W. (2010). Correlated expression patterns of microRNA genes with age-dependent behavioural changes in honey bee. *Insect Molecular*

Biology, *19*(4), 431–439. https://doi.org/10.1111/j.1365-2583.2010.01010.x

Benayoun, B. A., Pollina, E. A., & Brunet, A. (2015). Epigenetic regulation of ageing: Linking environmental inputs to genomic stability. *Nature Reviews-Molecular Cell Biology*, *16*(10), 593–610. https://doi.org/10.1038/nrm4048

Bewick, A. J., Vogel, K. J., Moore, A. J., & Schmitz, R. J. (2017). Evolution of DNA methylation across Insects. *Molecular Biology and Evolution*, *34*(3), 654–665. https://doi.org/10.1093/molbev/msw264

Beye, M., Härtel, S., Hagen, A., Hasselmann, M., & Omholt, S. W. (2002). Specific developmental gene silencing in the honey bee using a homeobox motif. *Insect Molecular Biology*, *11*(6), 527–532. https://doi.org/10.1046/j.1365-2583.2002.00361.x

Biggar, K. K., Charih, F., Liu, H., Ruiz-Blanco, Y. B., Stalker, L., Chopra, A., Connolly, J., Adhikary, H., Frensemier, K., Hoekstra, M., Galka, M., Fang, Q., Wynder, C., Stanford, W. L., Green, J. R., & Li, S. S.-C. (2020). Proteome-wide prediction of lysine methylation leads to identification of H2BK43 methylation and outlines the potential methyllysine proteome. *Cell Reports*, *32*(2), 107896. https://doi.org/10.1016/j.celrep.2020.107896

Blenau, W., & Baumann, A. (2016). Chapter 14 – octopaminergic and tyraminergic signaling in the honey bee (*Apis mellifera*) brain: Behavioral, pharmacological, and molecular aspects. In T. Farooqui & A. A. Farooqui (Eds.), *Trace amines and neurological disorders* (pp. 203–219). Academic Press.

Bobay, L.-M., Wissel, E. F., & Raymann, K. (2020). Strain structure and dynamics revealed by targeted deep sequencing of the honey bee gut microbiome. *mSphere*, *5*(4), e00694-20. https://doi.org/10.1128/msphere.00694-20

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., & SanCristobal, M. (2010). Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics*, *186*(1), 241–262. https://doi.org/10.1534/genetics.110.117275

Bonilla-Rosso, G., Steiner, T., Wichmann, F., Bexkens, E., & Engel, P. (2020). Honey bees harbor a diverse gut virome engaging in nested strain-level interactions with the microbiota. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(13), 7355–7362. https://doi.org/10.1073/pnas.2000228117

Bourgeois, Y. X. C., & Warren, B. H. (2021). An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes. *Molecular Ecology*, *30*(23), 6036–6071. https://doi.org/10.1111/mec.15989

Brascamp, E. W., & Bijma, P. (2014). Methods to estimate breeding values in honey bees. *Genetics, Selection, Evolution*, *46*(1), 53. https://doi.org/10.1186/s12711-014-0053-9

Brascamp, E. W., & Bijma, P. (2019). A note on genetic parameters and accuracy of estimated breeding values in honey bees. *Genetics, Selection, Evolution*, *51*(1), 71. https://doi.org/10.1186/s12711-019-0510-6

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Broadhurst, D., Goodacre, R., Reinke, S. N., Kuligowski, J., Wilson, I. D., Lewis, M. R., & Dunn, W. B. (2018). Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics*, *14*(6), 72. https://doi.org/10.1007/s11306-018-1367-3

Broadrup, R. L., Mayack, C., Schick, S. J., Eppley, E. J., White, H. K., & Macherone, A. (2019). Honey bee (*Apis mellifera*) exposomes and dysregulated metabolic pathways associated with *Nosema ceranae* infection. *PLOS One*, *14*(3), e0213249. https://doi.org/10.1371/journal.pone.0213249

Bromenshenk, J. J., Henderson, C. B., Wick, C. H., Stanford, M. F., Zulich, A. W., Jabbour, R. E., Deshpande, S. V., McCubbin, P. E., Seccomb, R. A., Welch, P. M., Williams, T., Firth, D. R., Skowronski, E., Lehmann, M. M., Bilimoria, S. L., Gress, J., Wanner, K. W., & Cramer, R. A. Jr. (2010). Iridovirus and microsporidian linked to honey bee colony decline. *PLOS One*, *5*(10), e13181. https://doi.org/10.1371/journal.pone.0013181

Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *American Journal of Human Genetics*, *108*(10), 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005

Brutscher, L. M., & Flenniken, M. L. (2015). RNAi and antiviral defense in the honey bee. *Journal of Immunology Research*, *2015*, 941897–941810. https://doi.org/10.1155/2015/941897

Burger, L., Gaidatzis, D., Schübeler, D., & Stadler, M. B. (2013). Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Research*, *41*(16), e155–e155. https://doi.org/10.1093/nar/gkt599

Buszewski, B., & Noga, S. (2012). Hydrophilic interaction liquid chromatography (HILIC)–a powerful separation technique. *Analytical and Bioanalytical Chemistry*, *402*(1), 231–247. https://doi.org/10.1007/s00216-011-5308-5

Calatayud-Vernich, P., Calatayud, F., Simó, E., & Picó, Y. (2018). Pesticide residues in honey bees, pollen and beeswax: Assessing beehive exposure. *Environmental Pollution*, *241*, 106–114. https://doi.org/10.1016/j.envpol.2018.05.062

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Cao, L., Zhao, X., Chen, Y., & Sun, C. (2021). Chromosome-scale genome assembly of the high royal jelly-producing honey bees. *Scientific Data*, *8*(1), 302. https://doi.org/10.1038/s41597-021-01091-7

Card, D. C., Shapiro, B., Giribet, G., Moritz, C., & Edwards, S. V. (2021). Museum genomics. *Annual Review of Genetics*, *55*(1), 633–659. https://doi.org/10.1146/annurev-genet-071719-020506

Carreck, N. L., Andree, M., Brent, C. S., Cox-Foster, D., Dade, H. A., Ellis, J. D., Hatjina, F., & van Englesdorp, D. (2013). Standard methods for *Apis mellifera* anatomy and dissection. In V. Dietemann, J. D. Ellis & P. Neumann (Eds.), *The COLOSS BEEBOOK. Volume I: Standard methods for Apis mellifera*. *Journal of Apicultural Research*, *52*(4), 1–40. https://doi.org/10.3896/IBRA.1.52.4.03

Carvalho, B. S., & Rustici, G. (2013). The challenges of delivering bioinformatics training in the analysis of high-throughput data. *Briefings in Bioinformatics*, *14*(5), 538–547. https://doi.org/10.1093/bib/bbt018

Castelli, L., Branchiccela, B., Garrido, M., Invernizzi, C., Porrini, M., Romero, H., Santos, E., Zunino, P., & Antúnez, K. (2020). Impact of nutritional stress on honey bee gut microbiota, immunity, and *Nosema ceranae* infection. *Microbial Ecology*, *80*(4), 908–919. https://doi.org/10.1007/s00248-020-01538-1

Chakrabarti, P., Morré, J. T., Lucas, H. M., Maier, C. S., & Sagili, R. R. (2019). The omics approach to bee nutritional landscape. *Metabolomics*, *15*(10), 127. https://doi.org/10.1007/s11306-019-1590-6

Chakravorty, S., Helb, D., Burday, M., Connell, N., & Alland, D. (2007). A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*, *69*(2), 330–339. https://doi.org/10.1016/j.mimet.2007.02.005

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & De Rosario, H. (2017). *Pwr: Basic functions for power analysis.* https://nyuscholars.nyu.edu/en/publications/pwr-basic-functions-for-power-analysis.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., De Rosario, H., De Rosario, M. H. (2018). *Package 'pwr.' R package Version* (vol. 1, issue no. 2). http://14.63.219.55/web/packages/pwr/pwr.pdf

Chandrasekaran, S., Rittschof, C. C., Djukovic, D., Gu, H., Raftery, D., Price, N. D., & Robinson, G. E. (2015). Aggression is associated with aerobic glycolysis in the honey bee brain. *Genes, Brain, and Behavior*, *14*(2), 158–166. https://doi.org/10.1111/gbb.12201

Chandrasekaran, S., Ament, S. A., Eddy, J. A., Rodriguez-Zas, S. L., Schatz, B. R., Price, N. D., & Robinson, G. E. (2011). Behavior-specific changes in transcriptional modules lead to distinct and predictable neurogenomic states. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(44), 18020–18025. https://doi.org/10.1073/pnas.1114093108

Chang, H., Ding, G., Jia, G., Feng, M., & Huang, J. (2022). Hemolymph metabolism analysis of honey bee (*Apis mellifera* L.) response to different bee pollens. *Insects*, *14*(1), 37. https://doi.org/10.3390/insects14010037

Chapman, N. C., Bourgeois, A. L., Beaman, L. D., Lim, J., Harpur, B. A., Zayed, A., Allsopp, M. H., Rinderer, T. E., & Oldroyd, B. P. (2017). An abbreviated SNP panel for ancestry assignment of honey bees (*Apis mellifera*). *Apidologie*, *48*(6), 776–783. https://doi.org/10.1007/s13592-017-0522-6

Chapman, N. C., Harpur, B. A., Lim, J., Rinderer, T. E., Allsopp, M. H., Zayed, A., & Oldroyd, B. P. (2015). A SNP test to identify Africanized honey bees via proportion of "African" ancestry. *Molecular Ecology Resources*, *15*(6), 1346–1355. https://doi.org/10.1111/1755-0998.12411

Chávez-Galarza, J., Henriques, D., Johnston, J. S., Azevedo, J. C., Patton, J. C., Muñoz, I., De la Rúa, P., & Pinto, M. A. (2013). Signatures of selection in the Iberian honey bee (*Apis mellifera iberiensis*) revealed by a genome scan analysis of single nucleotide polymorphisms. *Molecular Ecology*, *22*(23), 5890–5907. https://doi.org/10.1111/mec.12537

Chávez-Galarza, J., Henriques, D., Johnston, J. S., Carneiro, M., Rufino, J., Patton, J. C., & Pinto, M. A. (2015). Revisiting the Iberian honey bee (*Apis mellifera iberiensis*) contact zone: Maternal and genome-wide nuclear variations provide support for secondary contact from historical refugia. *Molecular Ecology*, *24*(12), 2973–2992. https://doi.org/10.1111/mec.13223

Chen, H., Wang, K., Ji, W., Xu, H., Liu, Y., Wang, S., Wang, Z., Gao, F., Lin, Z., & Ji, T. (2021). Metabolomic analysis of honey bees (*Apis mellifera*) response to carbendazim based on UPLC-MS. *Pesticide Biochemistry and Physiology*, *179*, 104975. https://doi.org/10.1016/j.pestbp.2021.104975

Chen, C., Liu, Z., Pan, Q., Chen, X., Wang, H., Guo, H., Liu, S., Lu, H., Tian, S., Li, R., & Shi, W. (2016). Genomic analyses reveal demographic history and temperate adaptation of the newly discovered honey bee subspecies *Apis mellifera* sinisxinyuan n. ssp. *Molecular Biology and Evolution*, *33*(5), 1337–1348. https://doi.org/10.1093/molbev/msw017

Chen, S., Hoene, M., Li, J., Li, Y., Zhao, X., Häring, H.-U., Schleicher, E. D., Weigert, C., Xu, G., & Lehmann, R. (2013). Simultaneous extraction of metabolome and lipidome with methyl tert-butyl ether from a single small tissue sample for ultra-high performance liquid chromatography/mass spectrometry. *Journal of Chromatography A*, *1298*, 9–16. https://doi.org/10.1016/j.chroma.2013.05.019

Chen, D., Sun, J., Zhu, J., Ding, X., Lan, T., Wang, X., Wu, W., Ou, Z., Zhu, L., Ding, P., Wang, H., Luo, L., Xiang, R., Wang, X., Qiu, J., Wang, S., Li, H., Chai, C., Liang, L., … Xu, X. (2021). Single cell atlas for 11 non-model mammals, reptiles and birds. *Nature Communications*, *12*(1), 7083. https://doi.org/10.1038/s41467-021-27162-2

Chen, Y., Li, E.-M., & Xu, L.-Y. (2022). Guide to metabolomics analysis: A bioinformatics workflow. *Metabolites*, *12*(4), 357. https://doi.org/10.3390/metabo12040357

Chen, Z., Traniello, I. M., Rana, S., Cash-Ahmed, A. C., Sankey, A. L., Yang, C., & Robinson, G. E. (2021). Neurodevelopmental and transcriptomic effects of CRISPR/Cas9-induced somatic orco mutation in honey bees. *Journal of Neurogenetics*, *35*(3), 320–332. https://doi.org/10.1080/01677063.2021.1887173

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

Chong, J., & Xia, J. (2018). MetaboAnalystR: An R package for flexible and reproducible analysis of metabolomics data. *Bioinformatics*, *34*(24), 4313–4314. https://doi.org/10.1093/bioinformatics/bty528

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: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., &

Tiedje, J. M. (2014). Ribosomal database project: Data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, *42*(Database issue), D633–D642. https://doi.org/10.1093/nar/gkt1244

Cologna, S. M., Russell, W. K., Lim, P. J., Vigh, G., & Russell, D. H. (2010). Combining isoelectric point-based fractionation, liquid chromatography and mass spectrometry to improve peptide detection and protein identification. *Journal of the American Society for Mass Spectrometry*, *21*(9), 1612–1619. https://doi.org/10.1016/j.jasms.2010.04.010

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Copeland, D. C., Maes, P. W., Mott, B. M., & Anderson, K. E. (2022). Changes in gut microbiota and metabolism associated with phenotypic plasticity in the honey bee *Apis mellifera*. *Frontiers in Microbiology*, *13*, 1059001. https://doi.org/10.3389/fmicb.2022.1059001

Cornman, S. R., Schatz, M. C., Johnston, S. J., Chen, Y.-P., Pettis, J., Hunt, G., Bourgeois, L., Elsik, C., Anderson, D., Grozinger, C. M., & Evans, J. D. (2010). Genomic survey of the ectoparasitic mite *Varroa destructor*, a major pest of the honey bee *Apis mellifera*. *BMC Genomics*, *11*(1), 602. https://doi.org/10.1186/1471-2164-11-602

Corona, M., Libbrecht, R., & Wheeler, D. E. (2016). Molecular mechanisms of phenotypic plasticity in social insects. *Current Opinion in Insect Science*, *13*, 55–60. https://doi.org/10.1016/j.cois.2015.12.003

Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, *26*(12), 1367–1372. https://doi.org/10.1038/nbt.1511

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., & Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Molecular & Cellular Proteomics*, *13*(9), 2513–2526. https://doi.org/10.1074/mcp.M113.031591

Crailsheim, K., Brodschneider, R., Aupinel, P., Behrens, D., Genersch, E., Vollmann, J., & Riessberger-Gallé, U. (2013). Standard methods for artificial rearing of *Apis mellifera* larvae. In V. Dietemann, J. D. Ellis & P. Neumann (Eds), *The COLOSS BEEBOOK, Volume I: Standard methods for Apis mellifera research*. Journal of Apicultural Research, 52(1). https://doi.org/10.3896/IBRA.1.52.1.05

Cridland, J. M., Ramirez, S. R., Dean, C. A., Sciligo, A., & Tsutsui, N. D. (2018). Genome sequencing of museum specimens reveals rapid changes in the genetic composition of honey bees in California. *Genome Biology and Evolution*, *10*(2), 458–472. https://doi.org/10.1093/gbe/evy007

Cridland, J. M., Tsutsui, N. D., & Ramírez, S. R. (2017). The complex demographic history and evolutionary origin of the western honey bee, *Apis mellifera*. *Genome Biology and Evolution*, *9*(2), 457–472. https://doi.org/10.1093/gbe/evx009

d'Errico, F., Backwell, L., Villa, P., Degano, I., Lucejko, J. J., Bamford, M. K., Higham, T. F. G., Colombini, M. P., & Beaumont, P. B. (2012). Early evidence of San material culture represented by organic artifacts from Border Cave, South Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(33), 13214–13219. https://doi.org/10.1073/pnas.1204213109

Daca-Roszak, P., Pfeifer, A., Żebracka-Gala, J., Jarząb, B., Witt, M., & Ziętkiewicz, E. (2016). EurEAs_Gplex—a new SNaPshot assay for continental population discrimination and gender identification. *Forensic Science International. Genetics*, *20*, 89–100. https://doi.org/10.1016/j.fsigen.2015.10.004

Daisley, B. A., & Reid, G. (2021). BEExact: A metataxonomic database tool for high-resolution inference of bee-associated microbial communities. *mSystems*, *6*(2), e00082-21. https://doi.org/10.1128/msystems.00082-21

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R, 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

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). https://doi.org/10.1093/gigascience/giab008

David, A., & Rostkowski, P. (2020). Chapter 2 – analytical techniques in metabolomics. In D. Álvarez-Muñoz & M. Farré (Eds.), *Environmental metabolomics* (pp. 35–64). Elsevier.

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, *6*(1), 226. https://doi.org/10.1186/s40168-018-0605-2

De Marino, A., Mahmoud, A. A., Bose, M., Bircan, K. O., Terpolovsky, A., Bamunusinghe, V., Bohn, S., Khan, U., Novković, B., & Yazdi, P. G. (2022). A comparative analysis of current phasing and imputation software. *PLOS One*, *17*(10), e0260177. https://doi.org/10.1371/journal.pone.0260177

De Mita, S., Thuillet, A.-C., Gay, L., Ahmadi, N., Manel, S., Ronfort, J., & Vigouroux, Y. (2013). Detecting selection along environmental gradients: Analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, *22*(5), 1383–1399. https://doi.org/10.1111/mec.12182

De Souza, D. P. (2013). Detection of polar metabolites through the use of gas chromatography-mass spectrometry. *Methods in Molecular Biology*, *1055*, 29–37. https://doi.org/10.1007/978-1-62703-577-4_3

de Villemereuil, P., Frichot, É., Bazin, É., François, O., & Gaggiotti, O. E. (2014). Genome scan methods against more complex models: When and how much should we trust them? *Molecular Ecology*, *23*(8), 2006–2019. https://doi.org/10.1111/mec.12705

Deboutte, W., Beller, L., Yinda, C. K., Maes, P., de Graaf, D. C., & Matthijnssens, J. (2020). Honey-bee–associated prokaryotic viral communities reveal wide viral diversity and a profound metabolic coding potential. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(19), 10511–10519. https://doi.org/10.1073/pnas.1921859117

Decker, L. E., San Juan, P. A., Warren, M. L., Duckworth, C. E., Gao, C., & Fukami, T. (2023). Higher variability in fungi compared to bacteria in the foraging honey bee gut. *Microbial Ecology*, *85*(1), 330–334. https://doi.org/10.1007/s00248-021-01922-5

Defossez, E., Bourquin, J., von Reuss, S., Rasmann, S., & Glauser, G. (2023). Eight key rules for successful data-dependent acquisition in mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, *42*(1), 131–143. https://doi.org/10.1002/mas.21715

Değirmenci, L., Geiger, D., Ferreira, F. L. R., Keller, A., Krischke, B., Beye, M., Steffan-Dewenter, I., & Scheiner, R. (2020). CRISPR/Cas9 mediated mutations as a new tool for studying taste in honeybees. *Chemical Senses, 45*(8), 655–666. https://doi.org/10.1101/2020.03.26.009696

De-Kayne, R., Frei, D., Greenway, R., Mendes, S. L., Retel, C., & Feulner, P. G. D. (2021). Sequencing platform shifts provide opportunities but pose challenges for combining genomic data sets. *Molecular Ecology Resources*, *21*(3), 653–660. https://doi.org/10.1111/1755-0998.13309

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, *10*(1), 5436. https://doi.org/10.1038/s41467-019-13225-y

Delaneau, O., & Zagury, J. (2012). Allele identification in assembled genomic sequence datasets. In F. Pompanon, & A. Bonin (Eds.), *Data production and analysis in population genomics: Methods and protocols* (pp. 197–211). Humana Press.

Deligkaris, K. (2022). Wikis: A viable platform for supporting academic research operations. *SSRN Electronic Journal*, *2*(6). https://doi.org/10.2139/ssrn.4214772

Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S., & Ralser, M. (2020). DIA-NN: Neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods*, *17*(1), 41–44. https://doi.org/10.1038/s41592-019-0638-x

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., & Hahn, M. W. (2014). Extensive error in the number of genes inferred from draft genome assemblies. *PLOS Computational Biology*, *10*(12), e1003998. https://doi.org/10.1371/journal.pcbi.1003998

Dickman, M. J., Kucharski, R., Maleszka, R., & Hurd, P. J. (2013). Extensive histone post-translational modification in honey bees. *Insect Biochemistry and Molecular Biology*, *43*(2), 125–137. https://doi.org/10.1016/j.ibmb.2012.11.003

Disayathanoowat, T., Li, H., Supapimon, N., Suwannarach, N., Lumyong, S., Chantawannakul, P., & Guo, J. (2020). Different dynamics of bacterial and fungal communities in hive-stored bee bread and their possible roles: A case study from two commercial honey bees in China. *Microorganisms*, *8*(2), 264. https://doi.org/10.3390/microorganisms8020264

Dobin, A., & Gingeras, T. R. (2015). Mapping RNA-seq reads with STAR. *Current Protocols in Bioinformatics, 51*, 11.14.1–11.14.19.

Doerr, A. (2015). DIA mass spectrometry. *Nature Methods*, *12*(1), 35–35. https://doi.org/10.1038/nmeth.3234

Dogantzis, K. A., Tiwari, T., Conflitti, I. M., Dey, A., Patch, H. M., Muli, E. M., Garnery, L., Whitfield, C. W., Stolle, E., Alqarni, A. S., Allsopp, M. H., & Zayed, A. (2021). Thrice out of Asia and the adaptive radiation of the western honey bee. *Science Advances*, *7*(49), eabj2151. https://doi.org/10.1126/sciadv.abj2151

Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N., & Rechavi, G. (2013). Transcriptome-wide mapping of N6-methyladenosine by m6A-seq based on immunocapturing and massively parallel sequencing. *Nature Protocols*, *8*(1), 176–189. https://doi.org/10.1038/nprot.2012.148

Drewell, R. A., Bush, E. C., Remnant, E. J., Wong, G. T., Beeler, S. M., Stringham, J. L., Lim, J., & Oldroyd, B. P. (2014). The dynamic DNA methylation cycle from egg to sperm in the honey bee *Apis mellifera*. *Development*, *141*(13), 2702–2711. https://doi.org/10.1242/dev.110163

Du Rand, E. E., Human, H., Smit, S., Beukes, M., Apostolides, Z., Nicolson, S. W., & Pirk, C. W. W. (2017). Proteomic and metabolomic analysis reveals rapid and extensive nicotine detoxification ability in honey bee larvae. *Insect Biochemistry and Molecular Biology*, *82*, 41–51. https://doi.org/10.1016/j.ibmb.2017.01.011

Duruz, S., Sevane, N., Selmoni, O., Vajana, E., Leempoel, K., Stucki, S., Orozco-terWengel, P., Rochat, E., Dunner, S., Bruford, M. W., & Joost, S, CLIMGEN Consortium. (2019). Rapid identification and interpretation of gene-environment associations using the new R.SamBada landscape genomics pipeline. *Molecular Ecology Resources*, *19*(5), 1355–1365. https://doi.org/10.1111/1755-0998.13044

Edelmann, M. J. (2011). Strong cation exchange chromatography in analysis of posttranslational modifications: Innovations and perspectives. *Journal of Biomedicine & Biotechnology*, *2011*, 936508. https://doi.org/10.1155/2011/936508

Ejigu, G. F., & Jung, J. (2020). Review on the computational genome annotation of sequences obtained by next-generation sequencing. *Biology*, *9*(9), 295. https://doi.org/10.3390/biology9090295

Ellegaard, K. M., & Engel, P. (2019). Genomic diversity landscape of the honey bee gut microbiota. *Nature Communications*, *10*(1), 446. https://doi.org/10.1038/s41467-019-08303-0

Elsik, C. G., Tayal, A., Unni, D. R., Burns, G. W., & Hagen, D. E. (2018). Hymenoptera genome database: Using HymenopteraMine to enhance genomic studies of hymenopteran insects. In M. Kollmar (Ed.), Methods in molecular biology (vol. 1757, pp. 513–556). Springer New York.

Elsik, C. G., Worley, K. C., Bennett, A. K., Beye, M., Camara, F., Childers, C. P., de Graaf, D. C., Debyser, G., Deng, J., Devreese, B., Elhaik, E., Evans, J. D., Foster, L. J., Graur, D., Guigo, R., Hoff, K. J., Holder, M. E., Hudson, M. E., Hunt, G. J., … Gibbs, R. A., Honey Bee Genome Sequencing Consortium. (2014). Finding the missing honey bee genes: Lessons learned from a genome upgrade. *BMC Genomics*, *15*(1), 86. https://doi.org/10.1186/1471-2164-15-86

Emerman, A. B., Bowman, S. K., Barry, A., Henig, N., Patel, K. M., Gardner, A. F., & Hendrickson, C. L. (2017). NEBNext direct: A novel, rapid, hybridization-based approach for the capture and library conversion of genomic regions of interest. *Current Protocols in Molecular Biology*, *119*(1), 7.30.1–7.30.24. https://doi.org/10.1002/cpmb.39

Engel, P., & Moran, N. A. (2013). Functional and evolutionary insights into the simple yet specific gut microbiota of the honey bee from metagenomic analysis. *Gut Microbes*, *4*(1), 60–65. https://doi.org/10.4161/gmic.22517

Engel, P., James, R. R., Koga, R., Kwong, W. K., McFrederick, Q. S., & Moran, N. A. (2013). Standard methods for research on *Apis mellifera* gut symbionts. *Journal of Apicultural Research*, *52*(4), 1–24. https://doi.org/10.3896/IBRA.1.52.4.07

Engel, P., Kwong, W. K., McFrederick, Q., Anderson, K. E., Barribeau, S. M., Chandler, J. A., Cornman, R. S., Dainat, J., de Miranda, J. R., Doublet, V., Emery, O., Evans, J. D., Farinelli, L., Flenniken, M. L., Granberg, F., Grasis, J. A.,

Gauthier, L., Hayer, J., Koch, H., … Dainat, B. (2016). The bee microbiome: impact on bee health and model for evolution and ecology of host-microbe interactions. *mBio*, *7*(2), e02164-15–e02115. https://doi.org/10.1128/mBio.02164-15

Estaki, M., Jiang, L., Bokulich, N. A., McDonald, D., González, A., Kosciolek, T., Martino, C., Zhu, Q., Birmingham, A., Vázquez-Baeza, Y., Dillon, M. R., Bolyen, E., Caporaso, J. G., & Knight, R. (2020). QIIME 2 enables comprehensive end-to-end analysis of diverse microbiome data and comparative studies with publicly available data. *Current Protocols in Bioinformatics*, *70*(1), e100. https://doi.org/10.1002/cpbi.100

Evans, J. D., Schwarz, R. S., Chen, Y.-P., Budge, G., Cornman, R. S., De La Rua, P., De Miranda, J. R., Foret, S., Foster, L., Gauthier, L., Genersch, E., Gisder, S., Jarosch, A., Kucharski, R., Lopez, D., Lun, C. M., Moritz, R. F. A., Maleszka, R., Muñoz, I., & Pinto, M. A. (2013). Standard methodologies for molecular research in *Apis mellifera*. In V. Dietemann, J. D. Ellis & P. Neumann (Eds.), *The COLOSS BEEBOOK, volume I: Standard methods for Apis mellifera research. Journal of Apicultural Research*, *52*(4), 1–54. https://doi.org/10.3896/IBRA.1.52.4.11

Everitt, T., Wallberg, A., Christmas, M. J., Olsson, A., Hoffmann, W., Neumann, P., & Webster, M. T. (2023). The genomic basis of adaptation to high elevations in Africanized honey bees. *Genome Biology and Evolution*, *15*(9), evad157. https://doi.org/10.1093/gbe/evad157

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*(3), 276–278. https://doi.org/10.1038/s41587-020-0439-x

Eynard, S. E., Vignal, A., Basso, B., Canale-Tabet, K., Le Conte, Y., Decourtye, A., Genestout, L., Labarthe, E., Mondet, F., & Servin, B. (2022). Reconstructing queen genotypes by pool sequencing colonies in eusocial insects: Statistical methods and their application to honey bee. *Molecular Ecology Resources*, *22*(8), 3035–3048. https://doi.org/10.1111/1755-0998.13685

Fang, F., Zhou, H., Feng, X., Chen, X., Wang, Z., Zhao, S., & Li, X. (2022). Gene expression and chromatin conformation differs between worker bees performing different tasks. *Genomics*, *114*(3), 110362. https://doi.org/10.1016/j.ygeno.2022.110362

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., & Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, *193*(3), 929–941. https://doi.org/10.1534/genetics.112.147231

Fauchery, L., Koriabine, M., Moore, L. P., Yoshinaga, Y., Barry, K., Kohler, A., & U'Ren, J. M. (2023). Tissue cultivation, preparation, and extraction of high molecular weight DNA for single-molecule genome sequencing of plant-associated fungi. In F. Martin & S. Uroz (Eds.), Methods in *molecular biology* (vol. 2605, pp. 79–102). Springer US.

Feinberg, A. P. (2007). Phenotypic plasticity and the epigenetics of human disease. *Nature*, *447*(7143), 433–440. https://doi.org/10.1038/nature05919

Feng, H., & Wu, H. (2019). Differential methylation analysis for bisulfite sequencing using DSS. *Quantitative Biology*, *7*(4), 327–334. https://doi.org/10.1007/s40484-019-0183-8

Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., Xie, D., Chen, G., Guo, C., Faircloth, B. C., Petersen, B., Wang, Z., Zhou, Q., Diekhans, M., Chen, W., Andreu-Sánchez, S., Margaryan, A., Howard, J. T., Parent, C., … Zhang, G. (2020). Dense sampling of bird diversity increases power of comparative genomics. *Nature*, *587*(7833), 252–257. https://doi.org/10.1038/s41586-020-2873-9

Fernández-Costa, C., Martínez-Bartolomé, S., McClatchy, D. B., Saviola, A. J., Yu, N.-K., & Yates, J. R., 3rd. (2020). Impact of the identification strategy on the reproducibility of the DDA and DIA results. *Journal of Proteome Research*, *19*(8), 3153–3161. https://doi.org/10.1021/acs.jproteome.0c00153

Flenniken, M. L., & Andino, R. (2013). Non-specific dsRNA-mediated antiviral response in the honey bee. *PLOS One*, *8*(10), e77263. https://doi.org/10.1371/journal.pone.0077263

Flesch, E. P., Rotella, J. J., Thomson, J. M., Graves, T. A., & Garrott, R. A. (2018). Evaluating sample size to estimate genetic management metrics in the genomics era. *Molecular Ecology Resources*, *18*(5), 1077–1091. https://doi.org/10.1111/1755-0998.12898

Foret, S., Kucharski, R., Pellegrini, M., Feng, S., Jacobsen, S. E., Robinson, G. E., & Maleszka, R. (2012). DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(13), 4968–4973. https://doi.org/10.1073/pnas.1202392109

Foster, S. D., Feutry, P., Grewe, P., & Davies, C. (2021). Sample size requirements for genetic studies on yellowfin tuna. *PLOS One*, *16*(11), e0259113. https://doi.org/10.1371/journal.pone.0259113

Foster, L. J. (2011). Interpretation of data underlying the link between colony collapse disorder (CCD) and an invertebrate iridescent virus. *Molecular & Cellular Proteomics*, *10*(3), M110.006387. https://doi.org/10.1074/mcp.M110.006387

Frau, A., Kenny, J. G., Lenzi, L., Campbell, B. J., Ijaz, U. Z., Duckworth, C. A., Burkitt, M. D., Hall, N., Anson, J., Darby, A. C., & Probert, C. S. J. (2019). DNA extraction and amplicon production strategies deeply influence the outcome of gut mycobiome studies. *Scientific Reports*, *9*(1), 9328. https://doi.org/10.1038/s41598-019-44974-x

Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, *30* (7), 1687–1699. https://doi.org/10.1093/molbev/mst063

Frith, M. C., & Kawaguchi, R. (2015). Split-alignment of genomes finds orthologies more accurately. *Genome Biology*, *16*(1), 106. https://doi.org/10.1186/s13059-015-0670-9

Fuller, Z. L., Niño, E. L., Patch, H. M., Bedoya-Reina, O. C., Baumgarten, T., Muli, E., Mumoki, F., Ratan, A., McGraw, J., Frazier, M., Masiga, D., Schuster, S., Grozinger, C. M., & Miller, W. (2015). Genome-wide analysis of signatures of selection in populations of African honey bees (*Apis mellifera*) using new web-based tools. *BMC Genomics*, *16*(1), 518. https://doi.org/10.1186/s12864-015-1712-0

Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M. (2024). BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research*, *34*(5), 769–777. https://doi.org/10.1101/gr.278090.123

Galbraith, D. A., Yang, X., Niño, E. L., Yi, S., & Grozinger, C. (2015). Parallel epigenomic and transcriptomic responses to viral infection in honey bees (*Apis mellifera*). *PLOS Pathogens*, *11*(3), e1004713. https://doi.org/10.1371/journal.ppat.1004713

Gallego Romero, I., Pai, A. A., Tung, J., & Gilad, Y. (2014). RNA-seq: Impact of RNA degradation on transcript quantification. *BMC Biology*, *12*(1), 42. https://doi.org/10.1186/1741-7007-12-42

Gillis, J., Mistry, M., & Pavlidis, P. (2010). Gene function analysis in complex data sets using ErmineJ. *Nature Protocols*, *5*(6), 1148–1159. https://doi.org/10.1038/nprot.2010.78

Glastad, K. M., Hunt, B. G., & Goodisman, M. A. (2014). Evolutionary insights into DNA methylation in insects. *Current Opinion in Insect Science*, *1*, 25–30. https://doi.org/10.1016/j.cois.2014.04.001

Glastad, K. M., Hunt, B. G., & Goodisman, M. A. D. (2019). Epigenetics in insects: Genome regulation and the generation of phenotypic diversity. *Annual Review of Entomology*, *64*(1), 185–203. https://doi.org/10.1146/annurev-ento-011118-111914

Gonçalves, R. S., & Musen, M. A. (2019). The variable quality of metadata about biological samples used in biomedical experiments. *Scientific Data*, *6*(1), 190021. https://doi.org/10.1038/sdata.2019.21

Grewe, F., Kronforst, M. R., Pierce, N. E., & Moreau, C. S. (2021). Museum genomics reveals the Xerces blue butterfly (*Glaucopsyche xerces*) was a distinct species driven to extinction. *Biology Letters*, *17*(7), 20210123. https://doi.org/10.1098/rsbl.2021.0123

Grozinger, C. M., & Flenniken, M. L. (2019). Bee viruses: Ecology, pathogenicity, and impacts. *Annual Review of Entomology*, *64*(1), 205–226. https://doi.org/10.1146/annurev-ento-011118-111942

Grozinger, C. M., & Robinson, G. E. (2015). The power and promise of applying genomics to honey bee health. *Current Opinion in Insect Science*, *10*, 124–132. https://doi.org/10.1016/j.cois.2015.03.007

Grozinger, C. M., & Zayed, A. (2020). Improving bee health through genomics. *Nature Reviews-Genetics*, *21*(5), 277–291. https://doi.org/10.1038/s41576-020-0216-1

Guarna, M. M., Hoover, S. E., Huxter, E., Higo, H., Moon, K.-M., Domanski, D., Bixby, M. E. F., Melathopoulos, A. P., Ibrahim, A., Peirson, M., Desai, S., Micholson, D., White, R., Borchers, C. H., Currie, R. W., Pernal, S. F., & Foster, L. J. (2017). Peptide biomarkers used for the selective breeding of a complex polygenic trait in honey bees. *Scientific Reports*, *7*(1), 8381. https://doi.org/10.1038/s41598-017-08464-2

Guichard, M., Dainat, B., Eynard, S., Vignal, A., Servin, B., & Neuditschko, M, Beestrong Consortium. (2021). Identification of quantitative trait loci associated with calmness and gentleness in honey bees using whole-genome sequences. *Animal Genetics*, *52*(4), 472–481. https://doi.org/10.1111/age.13070

Guijas, C., Montenegro-Burke, J. R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., Koellensperger, G., Huan, T., Uritboonthai, W., Aisporna, A. E., Wolan, D. W., Spilker, M. E., Benton, H. P., & Siuzdak, G. (2018). METLIN: A technology platform for identifying knowns and unknowns. *Analytical Chemistry*, *90*(5), 3156–3164. https://doi.org/10.1021/acs.analchem.7b04424

Günther, T., & Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics*, *195*(1), 205–220. https://doi.org/10.1534/genetics.113.152462

Guo, J., & Huan, T. (2020). Comparison of full-scan, data-dependent, and data-independent acquisition modes in liquid chromatography-mass spectrometry based untargeted metabolomics. *Analytical Chemistry*, *92*(12), 8072–8080. https://doi.org/10.1021/acs.analchem.9b05135

Guo, N., Zhao, L., Zhao, Y., Li, Q., Xue, X., Wu, L., Gomez Escalada, M., Wang, K., & Peng, W. (2020). Comparison of the chemical composition and biological activity of mature and immature honey: An HPLC/QTOF/MS-based metabolomic approach. *Journal of Agricultural and Food Chemistry*, *68*(13), 4062–4071. https://doi.org/10.1021/acs.jafc.9b07604

Haig, D. (2004). The (dual) origin of epigenetics. *Cold Spring Harbor Symposia on Quantitative Biology*, *69*(0), 67–70. https://doi.org/10.1101/sqb.2004.69.67

Hammer, T. J., & Moran, N. A. (2019). Links between metamorphosis and symbiosis in holometabolous insects. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *374*(1783), 20190068. https://doi.org/10.1098/rstb.2019.0068

Hammer, T. J., Dickerson, J. C., & Fierer, N. (2015). Evidence-based recommendations on storing and handling specimens for analyses of insect microbiota. *PeerJ*, *3*, e1190. https://doi.org/10.7717/peerj.1190

Hammer, T. J., Le, E., & Moran, N. A. (2021). Thermal niches of specialized gut symbionts: The case of social bees. *Proceedings-Biological Sciences*, *288*(1944), 20201480. https://doi.org/10.1098/rspb.2020.1480

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., … Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, *184*(13), 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

Harpur, B. A., Guarna, M. M., Huxter, E., Higo, H., Moon, K.-M., Hoover, S. E., Ibrahim, A., Melathopoulos, A. P., Desai, S., Currie, R. W., Pernal, S. F., Foster, L. J., & Zayed, A. (2019). Integrative genomics reveals the genetics and evolution of the honey bee's social immune system. *Genome Biology and Evolution*, *11*(3), 937–948. https://doi.org/10.1093/gbe/evz018

Harpur, B. A., Kadri, S. M., Orsi, R. O., Whitfield, C. W., & Zayed, A. (2020). Defense response in Brazilian honey bees (*Apis mellifera* scutellata × spp.) is underpinned by complex patterns of admixture. *Genome Biology and Evolution*, *12*(8), 1367–1377. https://doi.org/10.1093/gbe/evaa128

Harpur, B. A., Kent, C. F., Molodtsova, D., Lebon, J. M. D., Alqarni, A. S., Owayss, A. A., & Zayed, A. (2014). Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(7), 2614–2619. https://doi.org/10.1073/pnas.1315506111

Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*. https://search.proquest.com/openview/bc77cca0fb9390b44b9ef572fb574322/1?pq-origsite=gscholar&cbl=18750&casa_token=20UO4NDF2_MAAAAA:BxpBuNK-n8w6rnyrV94l9CEt1fYvf-qYgR_eTR4AeJnPsJgiwPt1oAxht7mrqIBN45dpket6Gw.

Hassanyar, A. K., Nie, H., Li, Z., Lin, Y., Huang, J., Woldegiorgis, S. T., Hussain, M., Feng, W., Zhang, Z., Yu, K., & Su, S. (2023). Discovery of SNP molecular markers and candidate genes associated with sacbrood virus resistance in *Apis cerana cerana* larvae by whole-genome

resequencing. *International Journal of Molecular Sciences*, *24*(7), 6238. https://doi.org/10.3390/ijms24076238

Hayes, B., & Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. This article is one of a selection of papers from the conference "Exploiting genome-wide association in oilseed Brassicas: A model for genetic improvement of major OECD crops for sustainable farming. *Genome*, *53*(11), 876–883. https://doi.org/10.1139/G10-076

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Hedrick, V. E., LaLand, M. N., Nakayasu, E. S., & Paul, L. N. (2015). Digestion, purification, and enrichment of protein samples for mass spectrometry. *Current Protocols in Chemical Biology*, *7*(3), 201–222. https://doi.org/10.1002/9780470559277.ch140272

Henriques, D., Browne, K. A., Barnett, M. W., Parejo, M., Kryger, P., Freeman, T. C., Muñoz, I., Garnery, L., Highet, F., Jonhston, J. S., McCormack, G. P., & Pinto, M. A. (2018). High sample throughput genotyping for estimating C-lineage introgression in the dark honey bee: An accurate and cost-effective SNP-based tool. *Scientific Reports*, *8*(1), 8552. https://doi.org/10.1038/s41598-018-26932-1

Henriques, D., Lopes, A. R., Chejanovsky, N., Dalmon, A., Higes, M., Jabal-Uriel, C., Le Conte, Y., Reyes-Carreño, M., Soroker, V., Martín-Hernández, R., & Pinto, M. A. (2021). A SNP assay for assessing diversity in immune genes in the honey bee (*Apis mellifera* L.). *Scientific Reports*, *11*(1), 15317. https://doi.org/10.1038/s41598-021-94833-x

Henriques, D., Parejo, M., Vignal, A., Wragg, D., Wallberg, A., Webster, M. T., & Pinto, M. A. (2018). Developing reduced SNP assays from whole-genome sequence data to estimate introgression in an organism with complex genetic patterns, the Iberian honey bee (*Apis mellifera iberiensis*). *Evolutionary Applications*, *11*(8), 1270–1282. https://doi.org/10.1111/eva.12623

Henriques, D., Wallberg, A., Chávez-Galarza, J., Johnston, J. S., Webster, M. T., & Pinto, M. A. (2018). Whole genome SNP-associated signatures of local adaptation in honey bees of the Iberian Peninsula. *Scientific Reports*, *8*(1), 11145. https://doi.org/10.1038/s41598-018-29469-5

Herb, B. R., Wolschin, F., Hansen, K. D., Aryee, M. J., Langmead, B., Irizarry, R., Amdam, G. V., & Feinberg, A. P. (2012). Reversible switching between epigenetic states in honey bee behavioral subcastes. *Nature Neuroscience*, *15*(10), 1371–1373. https://doi.org/10.1038/nn.3218

Hoencamp, C., Dudchenko, O., Elbatsh, A. M. O., Brahmachari, S., Raaijmakers, J. A., van Schaik, T., Sedeño Cacciatore, Á., Contessoto, V. G., van Heesbeen, R. G. H. P., van den Broek, B., Mhaskar, A. N., Teunissen, H., St Hilaire, B. G., Weisz, D., Omer, A. D., Pham, M., Colaric, Z., Yang, Z., Rao, S. S. P., … Rowland, B. D. (2021). 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science*, *372*(6545), 984–989. https://doi.org/10.1126/science.abe2218

Holman, L., Helanterä, H., Trontti, K., & Mikheyev, A. S. (2019). Comparative transcriptomics of social insect queen pheromones. *Nature Communications*, *10*(1), 1593. https://doi.org/10.1038/s41467-019-09567-2

Honey bee Genome Sequencing Consortium. (2006). Insights into social insects from the genome of the honey bee *Apis mellifera*. *Nature*, *443*(7114), 931–949. https://doi.org/10.1038/nature05260

Hortin, G. L., & Sviridov, D. (2010). The dynamic range problem in the analysis of the plasma proteome. *Journal of Proteomics*, *73*(3), 629–636. https://doi.org/10.1016/j.jprot.2009.07.001

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLOS Genetics*, *5*(6), e1000529. https://doi.org/10.1371/journal.pgen.1000529

Hroncova, Z., Havlik, J., Killer, J., Doskocil, I., Tyl, J., Kamler, M., Titera, D., Hakl, J., Mrazek, J., Bunesova, V., & Rada, V. (2015). Variation in honey bee gut microbial diversity affected by ontogenetic stage, age and geographic location. *PLOS One*, *10*(3), e0118707. https://doi.org/10.1371/journal.pone.0118707

Hu, M., Yu, J., Taylor, J. M. G., Chinnaiyan, A. M., & Qin, Z. S. (2010). On the detection and refinement of transcription factor binding sites using ChIP-Seq data. *Nucleic Acids Research*, *38*(7), 2154–2167. https://doi.org/10.1093/nar/gkp1180

Hu, X. F., Zhang, B., Liao, C. H., & Zeng, Z. J. (2019). High-efficiency CRISPR/Cas9-mediated gene editing in honey bee (*Apis mellifera*) embryos. *G3*, *9*(5), 1759–1766. https://doi.org/10.1534/g3.119.400130

Huang, H., Lin, S., Garcia, B. A., & Zhao, Y. (2015). Quantitative proteomic analysis of histone modifications. *Chemical Reviews*, *115*(6), 2376–2418. https://doi.org/10.1021/cr500491u

Huvenne, H., & Smagghe, G. (2010). Mechanisms of dsRNA uptake in insects and potential of RNAi for pest control: A review. *Journal of Insect Physiology*, *56*(3), 227–235. https://doi.org/10.1016/j.jinsphys.2009.10.004

Jakobsson, M., & Rosenberg, N. A. (2007). CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, *23*(14), 1801–1806. https://doi.org/10.1093/bioinformatics/btm233

Jarosch, A., & Moritz, R. F. A. (2012). RNA interference in honey bees: Off-target effects caused by dsRNA. *Apidologie*, *43*(2), 128–138. https://doi.org/10.1007/s13592-011-0092-y

Jin, M. J., Wang, Z. L., Wu, Z. H., He, X. J., Zhang, Y., Huang, Q., Zhang, L. Z., Wu, X. B., Yan, W. Y., & Zeng, Z. J. (2023). Phenotypic dimorphism between honey bee queen and worker is regulated by complicated epigenetic modifications. *iScience*, *26*(4), 106308. https://doi.org/10.1016/j.isci.2023.106308

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816–821. https://doi.org/10.1126/science.1225829

Jirtle, R. L., & Skinner, M. K. (2007). Environmental epigenomics and disease susceptibility. *Nature Reviews-Genetics*, *8*(4), 253–262. https://doi.org/10.1038/nrg2045

Jones, B. M., Rao, V. D., Gernat, T., Jagla, T., Cash-Ahmed, A. C., Rubin, B. E., Comi, T. J., Bhogale, S., Husain, S. S., Blatti, C., Middendorf, M., Sinha, S., Chandrasekaran, S., & Robinson, G. E. (2020). Individual differences in honey bee behavior enabled by plasticity in brain gene regulatory networks. *eLife*, *9*, 9. https://doi.org/10.7554/eLife.62850

Jones, J. C., Du, Z. G., Bernstein, R., Meyer, M., Hoppe, A., Schilling, E., Ableitner, M., Juling, K., Dick, R., Strauss,

A. S., & Bienefeld, K. (2020). Tool for genomic selection and breeding to evolutionary adaptation: Development of a 100K single nucleotide polymorphism array for the honey bee. *Ecology and Evolution*, *10*(13), 6246–6256. https://doi.org/10.1002/ece3.6357

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., Whibley, A., Becuwe, M., Baxter, S. W., Ferguson, L., Wilkinson, P. A., Salazar, C., Davidson, C., Clark, R., Quail, M. A., Beasley, H., Glithero, R., Lloyd, C., Sims, S., … Ffrench-Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, *477*(7363), 203–206. https://doi.org/10.1038/nature10341

Jourdan-Pineau, H., Antoine, G., Galataud, J., Delatte, H., Simiand, C., & Clémencet, J. (2021). Estimating heritability in honey bees: Comparison of three major methods based on empirical and simulated datasets. *Ecology and Evolution*, *11*(13), 8475–8486. https://doi.org/10.1002/ece3.7389

Jousse, C., Dalle, C., Abila, A., Traikia, M., Diogon, M., Lyan, B., El Alaoui, H., Vidau, C., & Delbac, F. (2020). A combined LC-MS and NMR approach to reveal metabolic changes in the hemolymph of honey bees infected by the gut parasite *Nosema ceranae*. *Journal of Invertebrate Pathology*, *176*, 107478. https://doi.org/10.1016/j.jip.2020.107478

Jung, J. (2023). Metabolomic Studies in *Apis mellifera*. *Journal of Apiculture*, *38*(2), 151–162. https://doi.org/10.17519/apiculture.2023.06.38.2.151

Kadlečková, D., Tachezy, R., Erban, T., Deboutte, W., Nunvář, J., Saláková, M., & Matthijnssens, J. (2022). The virome of healthy honey bee colonies: Ubiquitous occurrence of known and new viruses in bee populations. *mSystems*, *7*(3), e0007222. https://doi.org/10.1128/msystems.00072-22

Kakumanu, M. L., Reeves, A. M., Anderson, T. D., Rodrigues, R. R., & Williams, M. A. (2016). Honey bee gut microbiome is altered by in-hive pesticide exposures. *Frontiers in Microbiology*, *7*, 1255. https://doi.org/10.3389/fmicb.2016.01255

Kassambara, A., & Mundt, F. (2017). *Package 'factoextra.' Extract and visualize the results of multivariate data analyses* (vol. 76, no. 2). https://cran.microsoft.com/snapshot/2016-11-30/web/packages/factoextra/factoextra.pdf.

Keller, B. O., Sui, J., Young, A. B., & Whittal, R. M. (2008). Interferences and contaminants encountered in modern mass spectrometry. *Analytica Chimica Acta*, *627*(1), 71–81. https://doi.org/10.1016/j.aca.2008.04.043

Kešnerová, L., Mars, R. A. T., Ellegaard, K. M., Troilo, M., Sauer, U., & Engel, P. (2017). Disentangling metabolic functions of bacteria in the honey bee gut. *PLOS Biology*, *15*(12), e2003467. https://doi.org/10.1371/journal.pbio.2003467

Khan, K. A., Al-Ghamdi, A. A., Ghramh, H. A., Ansari, M. J., Ali, H., Alamri, S. A., Al-Kahtani, S. N., Adgaba, N., Qasim, M., & Hafeez, M. (2020). Structural diversity and functional variability of gut microbial communities associated with honey bees. *Microbial Pathogenesis*, *138*, 103793. https://doi.org/10.1016/j.micpath.2019.103793

Kille, B., Balaji, A., Sedlazeck, F. J., Nute, M., & Treangen, T. J. (2022). Multiple genome alignment in the telomere-to-telomere assembly era. *Genome Biology*, *23*(1), 182. https://doi.org/10.1186/s13059-022-02735-6

Kirkpatrick, M., & Barton, N. (2006). Chromosome inversions, local adaptation and speciation. *Genetics*, *173*(1), 419–434. https://doi.org/10.1534/genetics.105.047985

Kistler, L., Bieker, V. C., Martin, M. D., Pedersen, M. W., Ramos Madrigal, J., & Wales, N. (2020). Ancient plant genomics in archaeology, herbaria, and the environment. *Annual Review of Plant Biology*, *71*(1), 605–629. https://doi.org/10.1146/annurev-arplant-081519-035837

Klupczynska, A., Plewa, S., Dereziński, P., Garrett, T. J., Rubio, V. Y., Kokot, Z. J., & Matysiak, J. (2020). Identification and quantification of honey bee venom constituents by multiplatform metabolomics. *Scientific Reports*, *10*(1), 21645. https://doi.org/10.1038/s41598-020-78740-1

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C., & Schlötterer, C. (2011). PoPoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLOS One*, *6*(1), e15925. https://doi.org/10.1371/journal.pone.0015925

Kohno, H., & Kubo, T. (2018). mKast is dispensable for normal development and sexual maturation of the male European honey bee. *Scientific Reports*, *8*(1), 11877. https://doi.org/10.1038/s41598-018-30380-2

Kohno, H., & Kubo, T. (2019). Genetics in the honey bee: Achievements and prospects toward the functional analysis of molecular and neural mechanisms underlying social behaviors. *Insects*, *10*(10), 348. https://doi.org/10.3390/insects10100348

Kohno, H., Suenami, S., Takeuchi, H., Sasaki, T., & Kubo, T. (2016). Production of knockout mutants by CRISPR/Cas9 in the European honey bee, *Apis mellifera* L. *Zoological Science*, *33*(5), 505–512. https://doi.org/10.2108/zs160043

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. (2015). Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, *15*(5), 1179–1191. https://doi.org/10.1111/1755-0998.12387

Koulis, G. A., Tsagkaris, A. S., Aalizadeh, R., Dasenaki, M. E., Panagopoulou, E. I., Drivelos, S., Halagarda, M., Georgiou, C. A., Proestos, C., & Thomaidis, N. S. (2021). Honey phenolic compound profiling and authenticity assessment using HRMS targeted and untargeted metabolomics. *Molecules*, *26*(9), 2769. https://doi.org/10.3390/molecules26092769

Kowallik, V., & Mikheyev, A. S. (2021). Honey bee larval and adult microbiome life stages are effectively decoupled with vertical transmission overcoming early life perturbations. *mBio*, *12*(6), e0296621. https://doi.org/10.1128/mBio.02966-21

Kraus, F. B., Neumann, P., & Moritz, R. F. A. (2005). Genetic variance of mating frequency in the honey bee (*Apis mellifera* L.). *Insectes Sociaux*, *52*(1), 1–5. https://doi.org/10.1007/s00040-004-0766-9

Krimbas, C. B., & Powell, J. R. (1992). *Drosophila inversion polymorphism*. CRC Press.

Kristensen, A. R., Gsponer, J., & Foster, L. J. (2012). A high-throughput approach for measuring temporal changes in the interactome. *Nature Methods*, *9*(9), 907–909. https://doi.org/10.1038/nmeth.2131

Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, *27*(11), 1571–1572. https://doi.org/10.1093/bioinformatics/btr167

Kucharski, R., Maleszka, J., Foret, S., & Maleszka, R. (2008). Nutritional control of reproductive status in honey bees via DNA methylation. *Science*, *319*(5871), 1827–1830. https://doi.org/10.1126/science.1153069

Kwong, W. K., Mancenido, A. L., & Moran, N. A. (2017). Immune system stimulation by the native gut microbiota of honey bees. *Royal Society Open Science*, *4*(2), 170003. https://doi.org/10.1098/rsos.170003

Kwong, W. K., Medina, L. A., Koch, H., Sing, K.-W., Soh, E. J. Y., Ascher, J. S., Jaffé, R., & Moran, N. A. (2017). Dynamic microbiome evolution in social bees. *Science Advances*, *3*(3), e1600513. https://doi.org/10.1126/sciadv.1600513

LaFlamme, B. (2021). Genomes go platinum. *Nature Research*, https://doi.org/10.1038/d42859-020-00116-2

Lai, Z., Tsugawa, H., Wohlgemuth, G., Mehta, S., Mueller, M., Zheng, Y., Ogiwara, A., Meissen, J., Showalter, M., Takeuchi, K., Kind, T., Beal, P., Arita, M., & Fiehn, O. (2018). Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nature Methods*, *15*(1), 53–56. https://doi.org/10.1038/nmeth.4512

Landguth, E. L., Fedy, B. C., Oyler-McCANCE, S. J., Garey, A. L., Emel, S. L., Mumma, M., Wagner, H. H., Fortin, M.-J., & Cushman, S. A. (2012). Effects of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. *Molecular Ecology Resources*, *12*(2), 276–284. https://doi.org/10.1111/j.1755-0998.2011.03077.x

Langfelder, P., & Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, *31*(9), 814–821. https://doi.org/10.1038/nbt.2676

Lariviere, P. J., Leonard, S. P., Horak, R. D., Powell, J. E., & Barrick, J. E. (2023). Honey bee functional genomics using symbiont-mediated RNAi. *Nature Protocols*, *18*(3), 902–928. https://doi.org/10.1038/s41596-022-00778-4

Laukens, K., Naulaerts, S., & Berghe, W. V. (2015). Bioinformatics approaches for the functional interpretation of protein lists: From ontology term enrichment to network analysis. *Proteomics*, *15*(5–6), 981–996. https://doi.org/10.1002/pmic.201400296

Lawniczak, M. K. N., Durbin, R., Flicek, P., Lindblad-Toh, K., Wei, X., Archibald, J. M., Baker, W. J., Belov, K., Blaxter, M. L., Marques Bonet, T., Childers, A. K., Coddington, J. A., Crandall, K. A., Crawford, A. J., Davey, R. P., Di Palma, F., Fang, Q., Haerty, W., Hall, N., … Richards, S. (2022). Standards recommendations for the earth BioGenome project. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4), e2115639118. https://doi.org/10.1073/pnas.2115639118

Lee, H. K., Braynen, W., Keshav, K., & Pavlidis, P. (2005). ErmineJ: Tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, *6*(1), 269. https://doi.org/10.1186/1471-2105-6-269

Lee, J., & Lee, S. H. (2019). Development of a film-assisted honey bee egg collection system (FECS). *Apidologie*, *50*(6), 804–810. https://doi.org/10.1007/s13592-019-00687-8

Leonard, S. P., Powell, J. E., Perutka, J., Geng, P., Heckmann, L. C., Horak, R. D., Davies, B. W., Ellington, A. D., Barrick, J. E., & Moran, N. A. (2020). Engineered symbionts activate honey bee immunity and limit pathogens. *Science*, *367*(6477), 573–576. https://doi.org/10.1126/science.aax9039

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

Li, H., Janssens, J., De Waegeneer, M., Kolluru, S. S., Davie, K., Gardeux, V., Saelens, W., David, F. P. A., Brbić, M., Spanier, K., Leskovec, J., McLaughlin, C. N., Xie, Q., Jones, R. C., Brueckner, K., Shim, J., Tattikota, S. G., Schnorrer, F., Rust, K., … Zinzen, R. P., FCA Consortium§. (2022). Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science*, *375*(6584), eabk2432. https://doi.org/10.1126/science.abk2432

Li, H., Qu, W., Obrycki, J. J., Meng, L., Zhou, X., Chu, D., & Li, B. (2020). Optimizing sample size for population genomic study in a global invasive lady beetle, *Harmonia Axyridis*. *Insects*, *11*(5), 290. https://doi.org/10.3390/insects11050290

Li, Q., Sun, M., Wan, Z., Liang, J., Betti, M., Hrynets, Y., Xue, X., Wu, L., & Wang, K. (2019). Bee pollen extracts modulate serum metabolism in lipopolysaccharide-induced acute lung injury mice with anti-inflammatory effects. *Journal of Agricultural and Food Chemistry*, *67*(28), 7855–7868. https://doi.org/10.1021/acs.jafc.9b03082

Li, Q., Li, N., Hu, X., Li, J., Du, Z., Chen, L., Yin, G., Duan, J., Zhang, H., Zhao, Y., Wang, J., & Li, N. (2011). Genome-wide mapping of DNA methylation in chicken. *PLOS One*, *6*(5), e19428. https://doi.org/10.1371/journal.pone.0019428

Li, Q., Wang, M., Zhang, P., Liu, Y., Guo, Q., Zhu, Y., Wen, T., Dai, X., Zhang, X., Nagel, M., Dethlefsen, B. H., Xie, N., Zhao, J., Jiang, W., Han, L., Wu, L., Zhong, W., Wang, Z., Wei, X., … Liu, W. (2022). A single-cell transcriptomic atlas tracking the neural basis of division of labour in an ant superorganism. *Nature Ecology & Evolution*, *6*(8), 1191–1204. https://doi.org/10.1038/s41559-022-01784-1

Li, R., Li, L., Xu, Y., & Yang, J. (2022). Machine learning meets omics: Applications and perspectives. *Briefings in Bioinformatics*, *23*(1), 1–22. https://doi.org/10.1093/bib/bbab460

Li, Y., Ge, X., Peng, F., Li, W., & Li, J. J. (2022). Exaggerated false positives by popular differential expression methods when analyzing human population samples. *Genome Biology*, *23*(1), 79. https://doi.org/10.1186/s13059-022-02648-4

Li, Z., Hou, M., Qiu, Y., Zhao, B., Nie, H., & Su, S. (2020). Changes in antioxidant enzymes activity and metabolomic profiles in the guts of honey bee (*Apis mellifera*) larvae infected with *Ascosphaera apis*. *Insects*, *11*(7), 419. https://doi.org/10.3390/insects11070419

Li Z.. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. In arXiv [q-bio.GN]. arXiv. http://arxiv.org/abs/1303.3997.

Liao, Y., Smyth, G. K., & Shi, W. (2013). The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, *41*(10), e108–e108. https://doi.org/10.1093/nar/gkt214

Liberti, J., Kay, T., Quinn, A., Kesner, L., Frank, E. T., Cabirol, A., Richardson, T. O., Engel, P., & Keller, L. (2022). The gut microbiota affects the social network of honey bees. *Nature Ecology & Evolution*, *6*(10), 1471–1479. https://doi.org/10.1038/s41559-022-01840-w

Li-Byarlay, H. (2016). The function of DNA methylation marks in social insects. *Frontiers in Ecology and Evolution*, *4*, 57. https://doi.org/10.3389/fevo.2016.00057

Li-Byarlay, H., Boncristiani, H., Howell, G., Herman, J., Clark, L., Strand, M. K., Tarpy, D., & Rueppell, O. (2020). Transcriptomic and epigenomic dynamics of honey bees in response to lethal viral infection. *Frontiers in Genetics*, *11*, 566320. https://doi.org/10.3389/fgene.2020.566320

Li-Byarlay, H., Li, Y., Stroud, H., Feng, S., Newman, T. C., Kaneda, M., Hou, K. K., Worley, K. C., Elsik, C. G., Wickline, S. A., Jacobsen, S. E., Ma, J., & Robinson, G. E. (2013). RNA interference knockdown of DNA methyltransferase 3 affects gene alternative splicing in the honey bee. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(31), 12750–12755. https://doi.org/10.1073/pnas.1310735110

Li-Byarlay, H., Rittschof, C. C., Massey, J. H., Pittendrigh, B. R., & Robinson, G. E. (2014). Socially responsive effects of brain oxidative metabolism on aggression. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(34), 12533–12537. https://doi.org/10.1073/pnas.1412306111

Liu, C., Wang, M., Wei, X., Wu, L., Xu, J., Dai, X., Xia, J., Cheng, M., Yuan, Y., Zhang, P., Li, J., Feng, T., Chen, A., Zhang, W., Chen, F., Shang, Z., Zhang, X., Peters, B. A., & Liu, L. (2019). An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Scientific Data*, *6*(1), 65. https://doi.org/10.1038/s41597-019-0071-0

Liu, S., Wang, Z., Zhu, R., Wang, F., Cheng, Y., & Liu, Y. (2021). Three differential expression analysis methods for RNA sequencing: Limma, EdgeR, DESeq2. *Journal of Visualized Experiments*, *175*(175). https://doi.org/10.3791/62528

Liu, Z., Wu, F., Li, F., & Wei, Y. (2023). Methionine can reduce the sublethal risk of Chlorantraniliprole to honey bees (*Apis mellifera* L.): Based on metabolomics analysis. *Ecotoxicology and Environmental Safety*, *268*, 115682. https://doi.org/10.1016/j.ecoenv.2023.115682

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, *24*(5), 1031–1046. https://doi.org/10.1111/mec.13100

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lowe, R., Wojciechowski, M., Ellis, N., & Hurd, P. J. (2022). Chromatin accessibility-based characterisation of brain gene regulatory networks in three distinct honey bee polyphenisms. *Nucleic Acids Research*, *50*(20), 11550–11562. https://doi.org/10.1093/nar/gkac992

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: A tutorial. *Molecular Systems Biology*, *15*(6), e8746. https://doi.org/10.15252/msb.20188746

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., & Maleszka, R. (2010). The honey bee epigenomes: Differential methylation of brain DNA in queens and workers. *PLOS Biology*, *8*(11), e1000506. https://doi.org/10.1371/journal.pbio.1000506

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., & Lajoie, G. (2003). PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, *17*(20), 2337–2342. https://doi.org/10.1002/rcm.1196

Ma, C., Shi, X., Chen, S., Han, J., Bai, H., Li, Z., Li-Byarlay, H., & Bai, L. (2024). Combined pesticides in field doses weaken honey bee (*Apis cerana* F.) flight ability and analyses of transcriptomics and metabolomics. *Pesticide Biochemistry and Physiology*, *201*, 105793. https://doi.org/10.1016/j.pestbp.2024.105793

Maleszka, R. (2008). Epigenetic integration of environmental and genomic signals in honey bees: The critical interplay of nutritional, brain and reproductive networks. *Epigenetics*, *3*(4), 188–192. https://doi.org/10.4161/epi.3.4.6697

Manel, S., Albert, C. H., & Yoccoz, N. G. (2012). Sampling in landscape genomics. *Methods in Molecular Biology*, *888*, 3–12. https://doi.org/10.1007/978-1-61779-870-2_1

Manel, S., Schwartz, M. K., Luikart, G., & Taberlet, P. (2003). Landscape genetics: Combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, *18*(4), 189–197. https://doi.org/10.1016/S0169-5347(03)00008-9

Maori, E., Garbian, Y., Kunik, V., Mozes-Koch, R., Malka, O., Kalev, H., Sabath, N., Sela, I., & Shafir, S. (2019). A Transmissible RNA pathway in honey bees. *Cell Reports*, *27*(7), 1949–1959.e6. https://doi.org/10.1016/j.celrep.2019.04.073

Maori, E., Navarro, I. C., Boncristiani, H., Seilly, D. J., Rudolph, K. L. M., Sapetschnig, A., Lin, C.-C., Ladbury, J. E., Evans, J. D., Heeney, J. L., & Miska, E. A. (2019). A secreted RNA binding protein forms RNA-stabilizing granules in the honey bee royal jelly. *Molecular Cell*, *74*(3), 598–608.e6. https://doi.org/10.1016/j.molcel.2019.03.010

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, *14*(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Matsumura, Y., To, T. K., Kunieda, T., Kohno, H., Kakutani, T., & Kubo, T. (2022). Mblk-1/E93, an ecdysone related-transcription factor, targets synaptic plasticity-related genes in the honey bee mushroom bodies. *Scientific Reports*, *12*(1), 21367. https://doi.org/10.1038/s41598-022-23329-z

McAfee, A., Chapman, A., Higo, H., Underwood, R., Milone, J., Foster, L. J., Guarna, M. M., Tarpy, D. R., & Pettis, J. S. (2020). Vulnerability of honey bee queens to heat-induced loss of fertility. *Nature Sustainability*, *3*(5), 367–376. https://doi.org/10.1038/s41893-020-0493-x

McAfee, A., Chapman, A., Iovinella, I., Gallagher-Kurtzke, Y., Collins, T. F., Higo, H., Madilao, L. L., Pelosi, P., & Foster, L. J. (2018). A death pheromone, oleic acid, triggers hygienic behavior in honey bees (*Apis mellifera* L.). *Scientific Reports*, *8*(1), 5719. https://doi.org/10.1038/s41598-018-24054-2

McAfee, A., Harpur, B. A., Michaud, S., Beavis, R. C., Kent, C. F., Zayed, A., & Foster, L. J. (2016). Toward an upgraded honey bee (*Apis mellifera* L.) genome annotation using proteogenomics. *Journal of Proteome Research*, *15*(2), 411–421. https://doi.org/10.1021/acs.jproteome.5b00589

McAfee, A., Magaña, A. A., Foster, L. J., & Hoover, S. E. (2024). Differences in honeybee queen pheromones revealed by LC-MS/MS: Reassessing the honest signal hypothesis. *iScience, 27*(10), 110906. https://doi.org/10.1101/2024.04.19.590367

McAfee, A., Tarpy, D. R., & Foster, L. J. (2021). Queen honey bees exhibit variable resilience to temperature stress. *PLOS One*, *16*(8), e0255381. https://doi.org/10.1371/journal.pone.0255381

McAfee, A., Li, J., & Otte, M. (2022). Honey bee genome editing. In *Transgenic insects: Techniques and applications* (pp. 359–374). CABI.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Meng, J., Lu, Z., Liu, H., Zhang, L., Zhang, S., Chen, Y., Rao, M. K., & Huang, Y. (2014). A protocol for RNA methylation differential analysis with MeRIP-Seq data and exomePeak R/Bioconductor package. *Methods*, *69*(3), 274–281. https://doi.org/10.1016/j.ymeth.2014.06.008

Mikheyev, A. S., Tin, M. M. Y., Arora, J., & Seeley, T. D. (2015). Museum samples reveal rapid evolution by wild honey bees exposed to a novel parasite. *Nature Communications*, *6*(1), 7991. https://doi.org/10.1038/ncomms8991

Mikheyev, A. S., Zwick, A., Magrath, M. J. L., Grau, M. L., Qiu, L., Su, Y. N., & Yeates, D. (2017). Museum genomics confirms that the Lord Howe Island stick insect survived extinction. *Current Biology*, *27*(20), 3157–3161.e4. https://doi.org/10.1016/j.cub.2017.08.058

Milone, J. P., Chakrabarti, P., Sagili, R. R., & Tarpy, D. R. (2021). Colony-level pesticide exposure affects honey bee (*Apis mellifera* L.) royal jelly production and nutritional composition. *Chemosphere*, *263*, 128183. https://doi.org/10.1016/j.chemosphere.2020.128183

Misra, P., Jadhav, A. R., & Bapat, S. A. (2022). Single-cell sequencing: A cutting edge tool in molecular medical research. *Medical Journal, Armed Forces India*, *78*(Suppl 1), S7–S13. https://doi.org/10.1016/j.mjafi.2022.08.006

Moffitt, J. R., Lundberg, E., & Heyn, H. (2022). The emerging landscape of spatial profiling technologies. *Nature Reviews-Genetics*, *23*(12), 741–759. https://doi.org/10.1038/s41576-022-00515-3

Momeni, J., Parejo, M., Nielsen, R. O., Langa, J., Montes, I., Papoutsis, L., Farajzadeh, L., Bendixen, C., Căuia, E., Charrière, J.-D., Coffey, M. F., Costa, C., Dall'Olio, R., De la Rúa, P., Drazic, M. M., Filipi, J., Galea, T., Golubovski, M., Gregorc, A., … Estonba, A. (2021). Authoritative subspecies diagnosis tool for European honey bees based on ancestry informative SNPs. *BMC Genomics*, *22*(1), 101. https://doi.org/10.1186/s12864-021-07379-7

Morfin, N., Fillier, T. A., Pham, T. H., Goodwin, P. H., Thomas, R. H., & Guzman-Novoa, E. (2022). First insights into the honey bee (*Apis mellifera*) brain lipidome and its neonicotinoid-induced alterations associated with reduced self-grooming behavior. *Journal of Advanced Research*, *37*, 75–89. https://doi.org/10.1016/j.jare.2021.08.007

Moritz, R., & Southwick, E. E. (2012). *Bees as superorganisms: An evolutionary reality*. Springer Science & Business Media.

Morris, A. P. (2011). Transethnic meta-analysis of genome-wide association studies. *Genetic Epidemiology*, *35*(8), 809–822. https://doi.org/10.1002/gepi.20630

Motta, E. V. S., & Moran, N. A. (2020). Impact of glyphosate on the honey bee gut microbiota: Effects of intensity, duration, and timing of exposure. *mSystems*, *5*(4), 10–1128. https://doi.org/10.1128/mSystems.00268-20

Motta, E. V. S., & Moran, N. A. (2024). The honey bee microbiota and its impact on health and disease. *Nature Reviews-Microbiology*, *22*(3), 122–137. https://doi.org/10.1038/s41579-023-00990-3

Munjal, Y., Tonk, R., & Sharma, R. (2022). Analytical techniques used in metabolomics: A review. *Systematic Reviews in Pharmacy*, *11*(24), 2297–2318. https://doi.org/10.31858/0975-8453.13.8.515-521

Muñoz, I., Henriques, D., Jara, L., Johnston, J. S., Chávez-Galarza, J., De La Rúa, P., & Pinto, M. A. (2017). SNPs selected by information content outperform randomly selected microsatellite loci for delineating genetic identification and introgression in the Endangered dark European honey bee (*Apis mellifera mellifera*). *Molecular Ecology Resources*, *17*(4), 783–795. https://doi.org/10.1111/1755-0998.12637

Muñoz, I., Henriques, D., Johnston, J. S., Chávez-Galarza, J., Kryger, P., & Pinto, M. A. (2015). Reduced SNP panels for genetic identification and introgression analysis in the dark honey bee (*Apis mellifera mellifera*). *PLOS One*, *10*(4), e0124365. https://doi.org/10.1371/journal.pone.0124365

Muñoz-Colmenero, M., Baroja-Careaga, I., Kovačić, M., Filipi, J., Puškadija, Z., Kezić, N., Estonba, A., Büchler, R., & Zarraonaindia, I. (2020). Differences in honey bee bacterial diversity and composition in agricultural and pristine environments – a field study. *Apidologie*, *51*(6), 1018–1037. https://doi.org/10.1007/s13592-020-00779-w

Mushegian, A. A., & Tougeron, K. (2019). Animal-microbe interactions in the context of diapause. *The Biological Bulletin*, *237*(2), 180–191. https://doi.org/10.1086/706078

Musich, R., Cadle-Davidson, L., & Osier, M. V. (2021). Comparison of short-read sequence aligners indicates strengths and weaknesses for biologists to consider. *Frontiers in Plant Science*, *12*, 657240. https://doi.org/10.3389/fpls.2021.657240

Nakato, R., & Sakata, T. (2021). Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods*, *187*, 44–53. https://doi.org/10.1016/j.ymeth.2020.03.005

Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. (2017). Minimum sample sizes for population genomics: An empirical study from an Amazonian plant species. *Molecular Ecology Resources*, *17*(6), 1136–1147. https://doi.org/10.1111/1755-0998.12654

Neary, J. L., & Carless, M. A. (2020). Chapter nine – methylated DNA immunoprecipitation sequencing (MeDIP-seq): Principles and applications. In T. Tollefsbol (Ed.), *Epigenetics methods* (vol. 18, pp. 157–179). Academic Press.

Nelson, R. M., Wallberg, A., Simões, Z. L. P., Lawson, D. J., & Webster, M. T. (2017). Genomewide analysis of admixture and adaptation in the Africanized honey bee. *Molecular Ecology*, *26*(14), 3603–3617. https://doi.org/10.1111/mec.14122

Newton, I. L. G., & Roeselers, G. (2012). The effect of training set on the classification of honey bee gut microbiota using the Naïve Bayesian Classifier. *BMC Microbiology*, *12*(1), 221. https://doi.org/10.1186/1471-2180-12-221

Nguyen, N. H., Song, Z., Bates, S. T., Branco, S., Tedersoo, L., Menke, J., Schilling, J. S., & Kennedy, P. G. (2016). FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. *Fungal Ecology*, *20*, 241–248. https://doi.org/10.1016/j.funeco.2015.06.006

Nguyen, P. N., & Rehan, S. M. (2022). The effects of urban land use gradients on wild bee microbiomes. *Frontiers in Microbiology*, *13*, 992660. https://doi.org/10.3389/fmicb.2022.992660

Nunes, F. M. F., Aleixo, A. C., Barchuk, A. R., Bomtorin, A. D., Grozinger, C. M., & Simões, Z. L. P. (2013). Non-target effects of green fluorescent protein (GFP)-derived double-stranded RNA (dsRNA-GFP) used in honey bee RNA interference (RNAi) assays. *Insects*, *4*(1), 90–103. https://doi.org/10.3390/insects4010090

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Oldroyd, B. P., & Yagound, B. (2021). The role of epigenetics, particularly DNA methylation, in the evolution of caste in insect societies. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *376*(1826), 20200115. https://doi.org/10.1098/rstb.2020.0115

Oppenheim, S., Cao, X., Rueppel, O., Krongdang, S., Phokasem, P., DeSalle, R., Goodwin, S., Xing, J., Chantawannakul, P., & Rosenfeld, J. A. (2020). Whole genome sequencing and assembly of the Asian honey bee *Apis dorsata*. *Genome Biology and Evolution*, *12*(1), 3677–3683. https://doi.org/10.1093/gbe/evz277

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., & Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews Methods Primers*, *1*(1), 14. https://doi.org/10.1038/s43586-020-00011-0

Otte, M., Netschitailo, O., Kaftanoglu, O., Wang, Y., Page, R. E., Jr., & Beye, M. (2018). Improving genetic transformation rates in honey bees. *Scientific Reports*, *8*(1), 16534. https://doi.org/10.1038/s41598-018-34724-w

Oyler-McCance, S. J., Fedy, B. C., & Landguth, E. L. (2013). Sample design effects in landscape genetics. *Conservation Genetics*, *14*(2), 275–285. https://doi.org/10.1007/s10592-012-0415-1

Palmer, K. A., & Oldroyd, B. P. (2000). Evolution of multiple mating in the genus *Apis*. *Apidologie*, *31*(2), 235–248. https://doi.org/10.1051/apido:2000119

Pang, Z., Lu, Y., Zhou, G., Hui, F., Xu, L., Viau, C., Spigelman, A. F., MacDonald, P. E., Wishart, D. S., Li, S., & Xia, J. (2024). MetaboAnalyst 6.0: Towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Research*, *52*(W1), W398–W406. https://doi.org/10.1093/nar/gkae253

Panziera, D., Requier, F., Chantawannakul, P., Pirk, C. W. W., & Blacquière, T. (2022). The diversity decline in wild and managed honey bee populations urges for an integrated conservation approach. *Frontiers in Ecology and Evolution*, *10*, 10. https://doi.org/10.3389/fevo.2022.767950

Paradis, E. (2010). Pegas: An R package for population genetics with an integrated-modular approach. *Bioinformatics*, *26*(3), 419–420. https://doi.org/10.1093/bioinformatics/btp696

Paradis, E. (2020). *Population genomics with R*. CRC Press.

Parejo, M., Wragg, D., Henriques, D., Vignal, A., & Neuditschko, M. (2017). Genome-wide scans between two honey bee populations reveal putative signatures of human-mediated selection. *Animal Genetics*, *48*(6), 704–707. https://doi.org/10.1111/age.12599

Parejo, M., Wragg, D., Gauthier, L., Vignal, A., Neumann, P., & Neuditschko, M. (2016). Using whole-genome sequence information to foster conservation efforts for the European dark honey bee, *Apis mellifera mellifera*. *Frontiers in Ecology and Evolution*, *4*, 140. https://doi.org/10.3389/fevo.2016.00140

Parejo, M., Wragg, D., Henriques, D., Charrière, J.-D., & Estonba, A. (2020). Digging into the genomic past of swiss honey bees by whole-genome sequencing museum specimens. *Genome Biology and Evolution*, *12*(12), 2535–2551. https://doi.org/10.1093/gbe/evaa188

Paris, L., Peghaire, E., Moné, A., Diogon, M., Debroas, D., Delbac, F., & El Alaoui, H. (2020). Honey bee gut microbiota dysbiosis in pesticide/parasite co-exposures is mainly induced by *Nosema ceranae*. *Journal of Invertebrate Pathology*, *172*, 107348. https://doi.org/10.1016/j.jip.2020.107348

Paten, A. M., Taylor, M. C., Williams, M. R., Warden, A. C., & Kumar, A. (2022). Chapter 24 – acute sublethal exposure to a neonicotinoid pesticide triggers a short-term metabolic response in honey bee larvae⋆. In D. J. Beale, K. E. Hillyer, A. C. Warden, & O. A. H. Jones (Eds.), *Applied environmental metabolomics* (pp. 359–376). Academic Press.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Payne, S. H. (2015). The utility of protein and mRNA correlation. *Trends in Biochemical Sciences*, *40*(1), 1–3. https://doi.org/10.1016/j.tibs.2014.10.010

Peer, E., Rechavi, G., & Dominissini, D. (2017). Epitranscriptomics: Regulation of mRNA metabolism through modifications. *Current Opinion in Chemical Biology*, *41*, 93–98. https://doi.org/10.1016/j.cbpa.2017.10.008

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Peters, T. J., Buckley, M. J., Chen, Y., Smyth, G. K., Goodnow, C. C., & Clark, S. J. (2021). Calling differentially methylated regions from whole genome bisulphite sequencing with DMRcate. *Nucleic Acids Research*, *49*(19), e109–e109. https://doi.org/10.1093/nar/gkab637

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., & Lercher, M. J. (2014). PopGenome: An efficient Swiss army knife for population genomic analyses in R. *Molecular Biology and Evolution*, *31*(7), 1929–1936. https://doi.org/10.1093/molbev/msu136

Pfennig, D. W., Wund, M. A., Snell-Rood, E. C., Cruickshank, T., Schlichting, C. D., & Moczek, A. P. (2010). Phenotypic plasticity's impacts on diversification and speciation. *Trends in Ecology & Evolution*, *25*(8), 459–467. https://doi.org/10.1016/j.tree.2010.05.006

Pino, L. K., Just, S. C., MacCoss, M. J., & Searle, B. C. (2020). Acquiring and analyzing data independent acquisition proteomics experiments without spectrum libraries. *Molecular & Cellular Proteomics*, *19*(7), 1088–1103. https://doi.org/10.1074/mcp.P119.001913

Powell, E., Ratnayeke, N., & Moran, N. A. (2016). Strain diversity and host specificity in a specialized gut symbiont of honey bees and bumble bees. *Molecular Ecology*, *25*(18), 4461–4471. https://doi.org/10.1111/mec.13787

Powell, J. E., Carver, Z., Leonard, S. P., & Moran, N. A. (2021). Field-realistic tylosin exposure impacts honey bee microbiota and pathogen susceptibility, which is ameliorated by native gut probiotics. *Microbiology*

Spectrum, *9*(1), e0010321. https://doi.org/10.1128/Spectrum.00103-21

Prasad, N., Tarikere, S., Khanale, D., Habib, F., & Shashidhara, L. S. (2016). A comparative genomic analysis of targets of Hox protein Ultrabithorax amongst distant insect species. *Scientific Reports*, *6*(1), 27885. https://doi.org/10.1038/srep27885

Pratanwanich, P. N., Yao, F., Chen, Y., Koh, C. W. Q., Wan, Y. K., Hendra, C., Poon, P., Goh, Y. T., Yap, P. M. L., Chooi, J. Y., Chng, W. J., Ng, S. B., Thiery, A., Goh, W. S. S., & Göke, J. (2021). Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nature Biotechnology*, *39*(11), 1394–1402. https://doi.org/10.1038/s41587-021-00949-w

Pratavieira, M., da Silva Menegasso, A. R., Roat, T., Malaspina, O., & Palma, M. S. (2020). *In situ* metabolomics of the honey bee brain: The metabolism of l-arginine through the polyamine pathway in the proboscis extension response (PER). *Journal of Proteome Research*, *19*(2), 832–844. https://doi.org/10.1021/acs.jproteome.9b00653

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

Promoting best practice in nucleotide sequence data sharing. (2020). *Scientific Data*, 7(1), 152.

Puechmaille, S. J. (2016). The program structure does not reliably recover the correct population structure when sampling is uneven: Subsampling and new estimators alleviate the problem. *Molecular Ecology Resources*, *16*(3), 608–627. https://doi.org/10.1111/1755-0998.12512

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Qi, D., Lu, M., Li, J., & Ma, C. (2023). Metabolomics reveals distinctive metabolic profiles and marker compounds of Camellia (*Camellia sinensis* L.) bee pollen. *Foods 12*(14), 2661. https://doi.org/10.3390/foods12142661

Qiu, X., Sun, W., McDonnell, C. M., Li-Byarlay, H., Steele, L. D., Wu, J., Xie, J., Muir, W. M., & Pittendrigh, B. R. (2013). Genome-wide analysis of genes associated with moderate and high DDT resistance in *Drosophila melanogaster*. *Pest Management Science*, *69*(8), 930–937. https://doi.org/10.1002/ps.3454

Quail, M. A., Swerdlow, H., & Turner, D. J. (2009). Improved protocols for the illumina genome analyzer sequencing system. In Editorial Board, Jonathan L. Haines. *Current Protocols in Human Genetics*, *62*(1), 18.2.1–18.2.27. https://doi.org/10.1002/0471142905.hg1802s62

Quinn, A., El Chazli, Y., Escrig, S., Daraspe, J., Neuschwander, N., McNally, A., Genoud, C., Meibom, A., & Engel, P. (2024). Host-derived organic acids enable gut colonization of the honey bee symbiont *Snodgrassella alvi*. *Nature Microbiology*, *9*(2), 477–489. https://doi.org/10.1038/s41564-023-01572-y

Radloff, S. E., Hepburn, C., Randall Hepburn, H., Fuchs, S., Hadisoesilo, S., Tan, K., Engel, M. S., & Kuznetsov, V. (2010). Population structure and classification of *Apis cerana*. *Apidologie*, *41*(6), 589–601. https://doi.org/10.1051/apido/2010008

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Rampler, E., Abiead, Y. E., Schoeny, H., Rusz, M., Hildebrand, F., Fitz, V., & Koellensperger, G. (2021). Recurrent topics in mass spectrometry-based metabolomics and lipidomics-standardization, coverage, and throughput. *Analytical Chemistry*, *93*(1), 519–545. https://doi.org/10.1021/acs.analchem.0c04698

Rand, E. E. d., Smit, S., Beukes, M., Apostolides, Z., Pirk, C. W. W., & Nicolson, S. W. (2015). Detoxification mechanisms of honey bees (*Apis mellifera*) resulting in tolerance of dietary nicotine. *Scientific Reports*, *5*(1), 11779. https://doi.org/10.1038/srep11779

Randall Hepburn, H., & Radloff, S. E. (2011). *Honeybees of Asia*. Springer Science & Business Media.

Rappsilber, J., Ishihama, Y., & Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Analytical Chemistry*, *75*(3), 663–670. https://doi.org/10.1021/ac026117i

Raxworthy, C. J., & Smith, B. T. (2021). Mining museums for historical DNA: Advances and challenges in museomics. *Trends in Ecology & Evolution*, *36*(11), 1049–1060. https://doi.org/10.1016/j.tree.2021.07.009

Raymann, K., Coon, K. L., Shaffer, Z., Salisbury, S., & Moran, N. A. (2018). Pathogenicity of *Serratia marcescens* strains in honey bees. *mBio*, *9*(5), 10–1128. https://doi.org/10.1128/mBio.01649-18

Rehan, S. M., & Toth, A. L. (2015). Climbing the social ladder: The molecular evolution of sociality. *Trends in Ecology & Evolution*, *30*(7), 426–433. https://doi.org/10.1016/j.tree.2015.05.004

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, *24*(17), 4348–4370. https://doi.org/10.1111/mec.13322

Remnant, E. J., Ashe, A., Young, P. E., Buchmann, G., Beekman, M., Allsopp, M. H., Suter, C. M., Drewell, R. A., & Oldroyd, B. P. (2016). Parent-of-origin effects on genome-wide DNA methylation in the Cape honey bee (*Apis mellifera capensis*) may be confounded by allele-specific methylation. *BMC Genomics*, *17*(1), 226. https://doi.org/10.1186/s12864-016-2506-8

Rice, E. S., & Green, R. E. (2019). New approaches for genome assembly and scaffolding. *Annual Review of Animal Biosciences*, *7*(1), 17–40. https://doi.org/10.1146/annurev-animal-020518-115344

Ricigliano, V. A., Cank, K. B., Todd, D. A., Knowles, S. L., & Oberlies, N. H. (2022). Metabolomics-guided comparison of pollen and microalgae-based artificial diets in honey bees. *Journal of Agricultural and Food Chemistry*, *70*(31), 9790–9801. https://doi.org/10.1021/acs.jafc.2c02583

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47–e47. https://doi.org/10.1093/nar/gkv007

Rittschof, C. C., Bukhari, S. A., Sloofman, L. G., Troy, J. M., Caetano-Anollés, D., Cash-Ahmed, A., Kent, M., Lu, X., Sanogo, Y. O., Weisner, P. A., Zhang, H., Bell, A. M., Ma, J., Sinha, S., Robinson, G. E., & Stubbs, L. (2014). Neuromolecular responses to social challenge: Common mechanisms across mouse, stickleback fish, and honey bee. *Proceedings of the National Academy of Sciences of*

the United States of America, *111*(50), 17929–17934. https://doi.org/10.1073/pnas.1420369111

Rizwan, M., Liang, P., Ali, H., Li, Z., Nie, H., Ahmed Saqib, H. S., Fiaz, S., Raza, M. F., Hassanyar, A. K., Niu, Q., & Su, S. (2020). Population genomics of honey bees reveals a selection signature indispensable for royal jelly production. *Molecular and Cellular Probes*, *52*, 101542. https://doi.org/10.1016/j.mcp.2020.101542

Roffet-Salque, M., Regert, M., Evershed, R. P., Outram, A. K., Cramp, L. J. E., Decavallas, O., Dunne, J., Gerbault, P., Mileto, S., Mirabaud, S., Pääkkönen, M., Smyth, J., Šoberl, L., Whelton, H. L., Alday-Ruiz, A., Asplund, H., Bartkowiak, M., Bayer-Niemeier, E., Belhouchet, L., … Zoughlami, J. (2015). Widespread exploitation of the honey bee by early Neolithic farmers. *Nature*, *527*(7577), 226–230. https://doi.org/10.1038/nature15757

Romero, S., Nastasa, A., Chapman, A., Kwong, W. K., & Foster, L. J. (2019). The honey bee gut microbiota: Strategies for study and characterization. *Insect Molecular Biology*, *28*(4), 455–472. https://doi.org/10.1111/imb.12567

Rosenberg, N. A. (2004). Distruct: A program for the graphical display of population structure. *Molecular Ecology Notes*, *4*(1), 137–138. https://doi.org/10.1046/j.1471-8286.2003.00566.x

Roth, A., Vleurinck, C., Netschitailo, O., Bauer, V., Otte, M., Kaftanoglu, O., Page, R. E., & Beye, M. (2019). A genetic switch for worker nutrition-mediated traits in honey bees. *PLOS Biology*, *17*(3), e3000171. https://doi.org/10.1371/journal.pbio.3000171

Rothman, J. A., Leger, L., Kirkwood, J. S., & McFrederick, Q. S. (2019). Cadmium and selenate exposure affects the honey bee microbiome and metabolome, and bee-associated bacteria show potential for bioaccumulation. *Applied and Environmental Microbiology*, *85*(21), e01411-19. https://doi.org/10.1128/AEM.01411-19

Roundtree, I. A., Evans, M. E., Pan, T., & He, C. (2017). Dynamic RNA modifications in gene expression regulation. *Cell*, *169*(7), 1187–1200. https://doi.org/10.1016/j.cell.2017.05.045

Ruden, D. M., Cingolani, P. E., Sen, A., Qu, W., Wang, L., Senut, M.-C., Garfinkel, M. D., Sollars, V. E., & Lu, X. (2015). Epigenetics as an answer to Darwin's "special difficulty," Part 2: Natural selection of metastable epialleles in honey bee castes. *Frontiers in Genetics*, *6*, 60. https://doi.org/10.3389/fgene.2015.00060

Ruffo, P., De Amicis, F., Giardina, E., & Conforti, F. L. (2023). Long-noncoding RNAs as epigenetic regulators in neurodegenerative diseases. *Neural Regeneration Research*, *18*(6), 1243–1248. https://doi.org/10.4103/1673-5374.358615

Saelao, P., Simone-Finstrom, M., Avalos, A., Bilodeau, L., Danka, R., de Guzman, L., Rinkevich, F., & Tokarz, P. (2020). Genome-wide patterns of differentiation within and among U.S. commercial honey bee stocks. *BMC Genomics*, *21*(1), 704. https://doi.org/10.1186/s12864-020-07111-x

Salmela, H., Harwood, G. P., Münch, D., Elsik, C. G., Herrero-Galán, E., Vartiainen, M. K., & Amdam, G. V. (2022). Nuclear translocation of vitellogenin in the honey bee (*Apis mellifera*). *Apidologie*, *53*(1), 13. https://doi.org/10.1007/s13592-022-00914-9

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*(1), 87. https://doi.org/10.1186/s12915-014-0087-z

Sánchez-Vásquez, E., Alata Jimenez, N., Vázquez, N. A., & Strobl-Mazzulla, P. H. (2018). Emerging role of dynamic RNA modifications during animal development. *Mechanisms of Development*, *154*, 24–32. https://doi.org/10.1016/j.mod.2018.04.002

Santos, P. K. F., Arias, M. C., & Kapheim, K. M. (2019). Loss of developmental diapause as prerequisite for social evolution in bees. *Biology Letters*, *15*(8), 20190398. https://doi.org/10.1098/rsbl.2019.0398

Savaryn, J. P., Toby, T. K., & Kelleher, N. L. (2016). A researcher's guide to mass spectrometry-based proteomics. *Proteomics*, *16*(18), 2435–2443. https://doi.org/10.1002/pmic.201600113

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, *78*(4), 629–644. https://doi.org/10.1086/502802

Schloss, P. D. (2020). Reintroducing mothur: 10 Years later. *Applied and Environmental Microbiology*, *86*(2), e02343-19. https://doi.org/10.1128/AEM.02343-19

Schmitz, R. J., Lewis, Z. A., & Goll, M. G. (2019). DNA methylation: Shared and divergent features across eukaryotes. *Trends in Genetics*, *35*(11), 818–827. https://doi.org/10.1016/j.tig.2019.07.007

Schulte, C., Theilenberg, E., Müller-Borg, M., Gempe, T., & Beye, M. (2014). Highly efficient integration and expression of piggyBac-derived cassettes in the honey bee (*Apis mellifera*). *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 9003–9008. https://doi.org/10.1073/pnas.1402341111

Schwartz, M. K., & McKelvey, K. S. (2009). Why sampling scheme matters: The effect of sampling scheme on landscape genetic results. *Conservation Genetics*, *10*(2), 441–452. https://doi.org/10.1007/s10592-008-9622-1

Shames, D. S., Minna, J. D., & Gazdar, A. F. (2007). DNA methylation in health, disease, and cancer. *Current Molecular Medicine*, *7*(1), 85–102. https://doi.org/10.2174/156652407779940413

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: Past, present and future. *Nature*, *550*(7676), 345–353. https://doi.org/10.1038/nature24286

Shi, Y. Y., Huang, Z. Y., Zeng, Z. J., Wang, Z. L., Wu, X. B., & Yan, W. Y. (2011). Diet and cell size both affect queen-worker differentiation through DNA methylation in honey bees (*Apis mellifera*, Apidae). *PLOS One*, *6*(4), e18808. https://doi.org/10.1371/journal.pone.0018808

Shi, P., Zhou, J., Song, H., Wu, Y., Lan, L., Tang, X., Ma, Z., Vossbrinck, C. R., Vossbrinck, B., Zhou, Z., & Xu, J. (2020). Genomic analysis of Asian honey bee populations in China reveals evolutionary relationships and adaptation to abiotic stress. *Ecology and Evolution*, *10*(23), 13427–13438. https://doi.org/10.1002/ece3.6946

Shi, T., Burton, S., Wang, Y., Xu, S., Zhang, W., & Yu, L. (2018). Metabolomic analysis of honey bee, *Apis mellifera* L. response to thiacloprid. *Pesticide Biochemistry and Physiology*, *152*, 17–23. https://doi.org/10.1016/j.pestbp.2018.08.003

Shi, Y. Y., Wu, X. B., Huang, Z. Y., Wang, Z. L., Yan, W. Y., & Zeng, Z. J. (2012). Epigenetic modification of gene expression in honey bees by heterospecific gland secretions. *PLOS One*, *7*(8), e43727. https://doi.org/10.1371/journal.pone.0043727

Shi, Y. Y., Yan, W. Y., Huang, Z. Y., Wang, Z. L., Wu, X. B., & Zeng, Z. J. (2013). Genomewide analysis indicates that queen larvae have lower methylation levels in the honey bee (*Apis mellifera*). *Die Naturwissenschaften*, *100*(2), 193–197. https://doi.org/10.1007/s00114-012-1004-3

Shirvaliloo, M. (2022). The landscape of histone modifications in epigenomics since 2020. *Epigenomics*, *14*(23), 1465–1477. https://doi.org/10.2217/epi-2022-0437

Short, A. E. Z., Dikow, T., & Moreau, C. S. (2018). entomological collections in the age of big data. *Annual Review of Entomology*, *63*(1), 513–530. https://doi.org/10.1146/annurev-ento-031616-035536

Shpigler, H. Y., Saul, M. C., Murdoch, E. E., Cash-Ahmed, A. C., Seward, C. H., Sloofman, L., Chandrasekaran, S., Sinha, S., Stubbs, L. J., & Robinson, G. E. (2017). Behavioral, transcriptomic and epigenetic responses to social challenge in honey bees. *Genes, Brain, and Behavior*, *16*(6), 579–591. https://doi.org/10.1111/gbb.12379

Shpigler, H. Y., Saul, M. C., Murdoch, E. E., Corona, F., Cash-Ahmed, A. C., Seward, C. H., Chandrasekaran, S., Stubbs, L. J., & Robinson, G. E. (2019). Honey bee neurogenomic responses to affiliative and agonistic social interactions. *Genes, Brain, and Behavior*, *18*(1), e12509. https://doi.org/10.1111/gbb.12509

Sinitcyn, P., Hamzeiy, H., Salinas Soto, F., Itzhak, D., McCarthy, F., Wichmann, C., Steger, M., Ohmayer, U., Distler, U., Kaspar-Schoenefeld, S., Prianichnikov, N., Yılmaz, Ş., Rudolph, J. D., Tenzer, S., Perez-Riverol, Y., Nagaraj, N., Humphrey, S. J., & Cox, J. (2021). MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nature Biotechnology*, *39*(12), 1563–1573. https://doi.org/10.1038/s41587-021-00968-7

Sitnikov, D. G., Monnin, C. S., & Vuckovic, D. (2016). Systematic assessment of seven solvent and solid-phase extraction methods for metabolomics analysis of human plasma by LC-MS. *Scientific Reports*, *6*(1), 38885. https://doi.org/10.1038/srep38885

Snart, C. J. P., Hardy, I. C. W., & Barrett, D. A. (2015). Entometabolomics: Applications of modern analytical techniques to insect studies. *Entomologia Experimentalis et Applicata*, *155*(1), 1–17. https://doi.org/10.1111/eea.12281

Sokolowski, M. B. (2020). Honey bee colony aggression and indirect genetic effects [Review of honey bee colony aggression and indirect genetic effects]. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(31), 18148–18150. https://doi.org/10.1073/pnas.2012366117

Song, Q., Decato, B., Hong, E. E., Zhou, M., Fang, F., Qu, J., Garvin, T., Kessler, M., Zhou, J., & Smith, A. D. (2013). A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLOS One*, *8*(12), e81148. https://doi.org/10.1371/journal.pone.0081148

Southey, B. R., Zhu, P., Carr-Markell, M. K., Liang, Z. S., Zayed, A., Li, R., Robinson, G. E., & Rodriguez-Zas, S. L. (2016). Characterization of genomic variants associated with scout and recruit behavioral castes in honey bees using whole-genome sequencing. *PLOS One*, *11*(1), e0146430. https://doi.org/10.1371/journal.pone.0146430

Spannhoff, A., Kim, Y. K., Raynal, N. J.-M., Gharibyan, V., Su, M.-B., Zhou, Y.-Y., Li, J., Castellano, S., Sbardella, G., Issa, J.-P. J., & Bedford, M. T. (2011). Histone deacetylase inhibitor activity in royal jelly might facilitate caste switching in bees. *EMBO Reports*, *12*(3), 238–243. https://doi.org/10.1038/embor.2011.9

Speed, D., Hemani, G., Johnson, M. R., & Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, *91*(6), 1011–1021. https://doi.org/10.1016/j.ajhg.2012.10.010

Spicer, R. A., Salek, R., & Steinbeck, C. (2017). A decade after the metabolomics standards initiative it's time for a revision [Review of a decade after the metabolomics standards initiative it's time for a revision]. *Scientific Data*, *4*(1), 170138. https://doi.org/10.1038/sdata.2017.138

Spitz, F., & Furlong, E. E. M. (2012). Transcription factors: From enhancer binding to developmental control. *Nature Reviews-Genetics*, *13*(9), 613–626. https://doi.org/10.1038/nrg3207

Spötter, A., Gupta, P., Mayer, M., Reinsch, N., & Bienefeld, K. (2016). Genome-wide association study of a Varroa-specific defense behavior in honeybees (*Apis mellifera*). *The Journal of Heredity*, *107*(3), 220–227. https://doi.org/10.1093/jhered/esw005

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., Levine, A. J., La Manno, G., Skinnider, M. A., & Courtine, G. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications*, *12*(1), 5692. https://doi.org/10.1038/s41467-021-25960-2

Stadhouders, R., Filion, G. J., & Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. *Nature*, *569*(7756), 345–354. https://doi.org/10.1038/s41586-019-1182-7

Stephens, M. (2017). False discovery rates: A new deal. *Biostatistics*, *18*(2), 275–294. https://doi.org/10.1093/biostatistics/kxw041

Stucki, S., Orozco-terWengel, P., Forester, B. R., Duruz, S., Colli, L., Masembe, C., Negrini, R., Landguth, E., Jones, M. R., Bruford, M. W., Taberlet, P., & Joost, S, NEXTGEN Consortium. (2017). High performance computation of landscape genomic models including local indicators of spatial association. *Molecular Ecology Resources*, *17*(5), 1072–1089. https://doi.org/10.1111/1755-0998.12629

Suchan, T., Kusliy, M. A., Khan, N., Chauvey, L., Tonasso-Calvière, L., Schiavinato, S., Southon, J., Keller, M., Kitagawa, K., Krause, J., Bessudnov, A. N., Bessudnov, A. A., Graphodatsky, A. S., Valenzuela-Lamas, S., Wilczyński, J., Pospuła, S., Tunia, K., Nowak, M., Moskal-delHoyo, M., … Orlando, L. (2022). Performance and automation of ancient DNA capture with RNA hyRAD probes. *Molecular Ecology Resources*, *22*(3), 891–907. https://doi.org/10.1111/1755-0998.13518

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., Hankemeier, T., Hardy, N., Harnly, J., Higashi, R., Kopka, J., Lane, A. N., Lindon, J. C., Marriott, P., Nicholls, A. W., … Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*, *3*(3), 211–221. https://doi.org/10.1007/s11306-007-0082-2

Sun, J., Zhao, H., Wu, F., Zhu, M., Zhang, Y., Cheng, N., Xue, X., Wu, L., & Cao, W. (2021). Molecular mechanism of mature honey formation by GC-MS- and LC-MS-based metabolomics. *Journal of Agricultural and Food Chemistry*, *69*(11), 3362–3370. https://doi.org/10.1021/acs.jafc.1c00318

Tadano, H., Yamazaki, Y., Takeuchi, H., & Kubo, T. (2009). Age- and division-of-labour-dependent differential expression of a novel non-coding RNA, Nb-1, in the brain of worker honey bees, *Apis mellifera* L. *Insect Molecular Biology*, *18*(6), 715–726. https://doi.org/10.1111/j.1365-2583.2009.00911.x

Tang, H., Peng, J., Wang, P., & Risch, N. J. (2005). Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, *28*(4), 289–301. https://doi.org/10.1002/gepi.20064

Tang, K., Thornton, K. R., & Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLOS Biology*, *5*(7), e171. https://doi.org/10.1371/journal.pbio.0050171

Tarpy, D. R., Nielsen, R., & Nielsen, D. I. (2004). A scientific note on the revised estimates of effective paternity frequency in *Apis*. *Insectes Sociaux*, *51*(2), 203–204. https://doi.org/10.1007/s00040-004-0734-4

Techer, M. A., Rane, R. V., Grau, M. L., Roberts, J. M. K., Sullivan, S. T., Liachko, I., Childers, A. K., Evans, J. D., & Mikheyev, A. S. (2019). Divergent evolutionary trajectories following speciation in two ectoparasitic honey bee mites. *Communications Biology*, *2*(1), 357. https://doi.org/10.1038/s42003-019-0606-0

The Honeybee Genome Sequencing Consortium. (2006). Insights into social insects from the genome of the honey bee *Apis mellifera*. *Nature*, *443*(7114), 931.

Thibaud-Nissen, F., Souvorov, A., Murphy, T., DiCuccio, M., & Kitts, P. (2013). *Eukaryotic genome annotation pipeline*. National Center for Biotechnology Information.

Thioulouse, J., Dray, S., Dufour, A.-B., Siberchicot, A., Jombart, T., & Pavoine, S. (2018). *Multivariate analysis of ecological data with ade4*. Springer New York.

Tihelka, E., Cai, C., Pisani, D., & Donoghue, P. C. J. (2020). Mitochondrial genomes illuminate the evolutionary history of the Western honey bee (*Apis mellifera*). *Scientific Reports*, *10*(1), 14515. https://doi.org/10.1038/s41598-020-71393-0

Timp, W., & Timp, G. (2020). Beyond mass spectrometry, the next step in proteomics. *Science Advances*, *6*(2), eaax8978. https://doi.org/10.1126/sciadv.aax8978

Tirado-Magallanes, R., Rebbani, K., Lim, R., Pradhan, S., & Benoukraf, T. (2017). Whole genome DNA methylation: Beyond genes silencing. *Oncotarget*, *8*(3), 5629–5637. https://doi.org/10.18632/oncotarget.13562

Tokarz, R., Firth, C., Street, C., Cox-Foster, D. L., & Lipkin, W. I. (2011). Lack of evidence for an association between Iridovirus and colony collapse disorder. *PLOS One*, *6*(6), e21844. https://doi.org/10.1371/journal.pone.0021844

Toth, A. L., & Zayed, A. (2021). The honey bee genome–what has it been good for? *Apidologie*, *52*(1), 45–62. https://doi.org/10.1007/s13592-020-00829-3

Traniello, I. M., Bukhari, S. A., Dibaeinia, P., Serrano, G., Avalos, A., Ahmed, A. C., Sankey, A. L., Hernaez, M., Sinha, S., Zhao, S. D., Catchen, J., & Robinson, G. E. (2023). Single-cell dissection of aggression in honey bee colonies. *Nature Ecology & Evolution*, *7*(8), 1232–1244. https://doi.org/10.1038/s41559-023-02090-0

Traniello, I. M., Bukhari, S. A., Kevill, J., Ahmed, A. C., Hamilton, A. R., Naeger, N. L., Schroeder, D. C., & Robinson, G. E. (2020). Meta-analysis of honey bee neurogenomic response links deformed wing virus type A to precocious behavioral maturation. *Scientific Reports*, *10*(1), 3101. https://doi.org/10.1038/s41598-020-59808-4

Tsompana, M., & Buck, M. J. (2014). Chromatin accessibility: A window into the genome. *Epigenetics & Chromatin*, *7*(1), 33. https://doi.org/10.1186/1756-8935-7-33

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., & Arita, M. (2015). MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods*, *12*(6), 523–526. https://doi.org/10.1038/nmeth.3393

Turner, B. M. (2008). *Chromatin and gene regulation: Molecular mechanisms in epigenetics*. John Wiley & Sons.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, *13*(9), 731–740. https://doi.org/10.1038/nmeth.3901

Urbut, S. M., Wang, G., Carbonetto, P., & Stephens, M. (2019). Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature Genetics*, *51*(1), 187–195. https://doi.org/10.1038/s41588-018-0268-8

van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J., & Lander, E. S. (2010). Hi-C: A method to study the three-dimensional architecture of genomes. *Journal of Visualized Experiments*, *39*(39). https://doi.org/10.3791/1869

Veenstra, T. D. (2012). Metabolomics: The final frontier? *Genome Medicine*, *4*(4), 40. https://doi.org/10.1186/gm339

Verheggen, K., Raeder, H., Berven, F. S., Martens, L., Barsnes, H., & Vaudel, M. (2020). Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrometry Reviews*, *39*(3), 292–306. https://doi.org/10.1002/mas.21543

Vernier, C. L., Chin, I. M., Adu-Oppong, B., Krupp, J. J., Levine, J., Dantas, G., & Ben-Shahar, Y. (2020). The gut microbiome defines social group membership in honey bee colonies. *Science Advances*, *6*(42), eabd3431. https://doi.org/10.1126/sciadv.abd3431

Virgiliou, C., Kanelis, D., Pina, A., Gika, H., Tananaki, C., Zotou, A., & Theodoridis, G. (2020). A targeted approach for studying the effect of sugar bee feeding on the metabolic profile of Royal Jelly. *Journal of Chromatography. A*, *1616*, 460783. https://doi.org/10.1016/j.chroma.2019.460783

von Thaden, A., Nowak, C., Tiesmeyer, A., Reiners, T. E., Alves, P. C., Lyons, L. A., Mattucci, F., Randi, E., Cragnolini, M., Galián, J., Hegyeli, Z., Kitchener, A. C., Lambinet, C., Lucas, J. M., Mölich, T., Ramos, L., Schockert, V., & Cocchiararo, B. (2020). Applying genomic data in wildlife monitoring: Development guidelines for genotyping degraded samples with reduced single nucleotide polymorphism panels. *Molecular Ecology Resources*, *20*(3), 662–680. https://doi.org/10.1111/1755-0998.13136

Wallberg, A., Bunikis, I., Pettersson, O. V., Mosbech, M.-B., Childers, A. K., Evans, J. D., Mikheyev, A. S., Robertson, H. M., Robinson, G. E., & Webster, M. T. (2019). A hybrid de novo genome assembly of the honey bee, *Apis mellifera*, with chromosome-length scaffolds. *BMC Genomics*, *20*(1), 275. https://doi.org/10.1186/s12864-019-5642-0

Wallberg, A., Han, F., Wellhagen, G., Dahle, B., Kawata, M., Haddad, N., Simões, Z. L. P., Allsopp, M. H., Kandemir, I., De la Rúa, P., Pirk, C. W., & Webster, M. T. (2014). A worldwide survey of genome sequence variation

provides insight into the evolutionary history of the honey bee *Apis mellifera*. *Nature Genetics*, *46*(10), 1081–1088. https://doi.org/10.1038/ng.3077

Wallberg, A., Schöning, C., Webster, M. T., & Hasselmann, M. (2017). Two extended haplotype blocks are associated with adaptation to high altitude habitats in East African honey bees. *PLOS Genetics*, *13*(5), e1006792. https://doi.org/10.1371/journal.pgen.1006792

Walsh, A. T., Triant, D. A., Le Tourneau, J. J., Shamimuzzaman, M., & Elsik, C. G. (2022). Hymenoptera genome database: New genomes and annotation datasets for improved go enrichment and orthologue analyses. *Nucleic Acids Research*, *50*(D1), D1032–D1039. https://doi.org/10.1093/nar/gkab1018

Wang B., & Li-Byarlay, H. (2015). Physiological and molecular mechanisms of nutrition in honey bees. *Advances in Insect Physiology*, *49*, 25–58.

Wang, B., Habermehl, C., & Jiang, L. (2022). Metabolomic analysis of honey bee (*Apis mellifera* L.) response to glyphosate exposure. *Molecular Omics*, *18*(7), 635–642. https://doi.org/10.1039/d2mo00046f

Wang, X., Lin, Y., Liang, L., Geng, H., Zhang, M., Nie, H., & Su, S. (2021). Transcriptional profiles of diploid mutant *Apis mellifera* embryos after knockout of csd by CRISPR/Cas9. *Insects*, *12*(8), 704. https://doi.org/10.3390/insects12080704

Wang, X., Li, Y., Chen, L., & Zhou, J. (2022). Analytical strategies for LC-MS-based untargeted and targeted metabolomics approaches reveal the entomological origins of honey. *Journal of Agricultural and Food Chemistry*, *70*(4), 1358–1366. https://doi.org/10.1021/acs.jafc.1c07153

Wang, M., Xiao, Y., Li, Y., Wang, X., Qi, S., Wang, Y., Zhao, L., Wang, K., Peng, W., Luo, G.-Z., Xue, X., Jia, G., & Wu, L. (2021). RNA m6A modification functions in larval development and caste differentiation in honey bee (*Apis mellifera*). *Cell Reports*, *34*(1), 108580. https://doi.org/10.1016/j.celrep.2020.108580

Wang, Y., Jorda, M., Jones, P. L., Maleszka, R., Ling, X., Robertson, H. M., Mizzen, C. A., Peinado, M. A., & Robinson, G. E. (2006). Functional CpG methylation system in a social insect. *Science*, *314*(5799), 645–647. https://doi.org/10.1126/science.1135213

Wang, Z.-L., Zhu, Y.-Q., Yan, Q., Yan, W.-Y., Zheng, H.-J., & Zeng, Z.-J. (2020). A chromosome-scale assembly of the Asian honey bee *Apis cerana* genome. *Frontiers in Genetics*, *11*, 279. https://doi.org/10.3389/fgene.2020.00279

Warner, M. R., Qiu, L., Holmes, M. J., Mikheyev, A. S., & Linksvayer, T. A. (2019). Convergent eusocial evolution is based on a shared reproductive groundplan plus lineage-specific plastic genes. *Nature Communications*, *10*(1), 2651. https://doi.org/10.1038/s41467-019-10546-w

Wattanachaiyingcharoen, W., Oldroyd, B. P., Wongsiri, S., Palmer, K., & Paar, J. (2003). A scientific note on the mating frequency of *Apis dorsata*. *Apidologie*, *34*(1), 85–86. https://doi.org/10.1051/apido:2002044

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution, 38*(6), 1358–1370. https://doi.org/10.2307/2408641

Wemheuer, F., Taylor, J. A., Daniel, R., Johnston, E., Meinicke, P., Thomas, T., & Wemheuer, B. (2020). Tax4Fun2: Prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environmental Microbiome*, *15*(1), 11. https://doi.org/10.1186/s40793-020-00358-7

Wen, Y., Wang, L., Jin, Y., Zhang, J., Su, L., Zhang, X., Zhou, J., & Li, Y. (2017). The microbial community dynamics during the Vitex honey ripening process in the honeycomb. *Frontiers in Microbiology*, *8*, 1649. https://doi.org/10.3389/fmicb.2017.01649

Williams, G. R., Alaux, C., Costa, C., Csáki, T., Doublet, V., Eisenhardt, D., Fries, I., Kuhn, R., Mcmahon, D. P., Medrzycki, P., Murray, T. E., Natsopoulou, M. E., Neumann, P., Oliver, R., Paxton, R. J., Pernal, S. F., Shutler, D., Tanner, G., van der Steen, J. J. F., & Brodschneider, R. (2013). Standard methods for maintaining adult *Apis mellifera* in cages under in vitro laboratory conditions. In V. Dietemann, J. D. Ellis & P. Neumann (Eds.), *The COLOSS BEEBOOK, volume I: Standard methods for Apis mellifera research*. *Journal of Apicultural Research*, *52*(1), 1–36. https://doi.org/10.3896/IBRA.1.52.1.04

Willing, E.-M., Dreyer, C., & van Oosterhout, C. (2012). Estimates of genetic differentiation measured by F(ST) do not necessarily require large sample sizes when using many SNP markers. *PLOS One*, *7*(8), e42649. https://doi.org/10.1371/journal.pone.0042649

Wilson, M. B., Spivak, M., Hegeman, A. D., Rendahl, A., & Cohen, J. D. (2013). Metabolomics reveals the origins of antimicrobial plant resins collected by honey bees. *PLOS One*, *8*(10), e77512. https://doi.org/10.1371/journal.pone.0077512

Wilson, R. C., & Doudna, J. A. (2013). Molecular mechanisms of RNA interference. *Annual Review of Biophysics*, *42*(1), 217–239. https://doi.org/10.1146/annurev-biophys-083012-130404

Wojciechowski, M., Lowe, R., Maleszka, J., Conn, D., Maleszka, R., & Hurd, P. J. (2018). Phenotypically distinct female castes in honey bees are defined by alternative chromatin states during larval development. *Genome Research*, *28*(10), 1532–1542. https://doi.org/10.1101/gr.236497.118

Wragg, D., Eynard, S. E., Basso, B., Canale-Tabet, K., Labarthe, E., Bouchez, O., Bienefeld, K., Bieńkowska, M., Costa, C., Gregorc, A., Kryger, P., Parejo, M.A., Pinto, M., Bidanel, J.-P., Servin, B., Le Conte, Y., & Vignal, A. (2021). Complex population structure and haplotype patterns in Western Europe honey bee from sequencing a large panel of haploid drones. *Molecular Ecology Resources, 22*(8), 3068–3086. https://doi.org/10.1101/2021.09.20.460798

Wragg, D., Marti-Marimon, M., Basso, B., Bidanel, J.-P., Labarthe, E., Bouchez, O., Le Conte, Y., & Vignal, A. (2016). Whole-genome resequencing of honey bee drones to detect genomic selection in a population managed for royal jelly. *Scientific Reports*, *6*(1), 27168. https://doi.org/10.1038/srep27168

Wratten, L., Wilm, A., & Göke, J. (2021). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods*, *18*(10), 1161–1168. https://doi.org/10.1038/s41592-021-01254-9

Wu, J., Lang, H., Mu, X., Zhang, Z., Su, Q., Hu, X., & Zheng, H. (2021). Honey bee genetics shape the strain-level structure of gut microbiota in social transmission. *Microbiome, 9*, 1–19. https://doi.org/10.1101/2020.12.17.423353

Wu, J.-L., Zhou, C.-X., Wu, P.-J., Xu, J., Guo, Y.-Q., Xue, F., Getachew, A., & Xu, S.-F. (2017). Brain metabolomic profiling of eastern honey bee (*Apis cerana*) infested with the mite *Varroa destructor*. *PLOS One*, *12*(4), e0175573. https://doi.org/10.1371/journal.pone.0175573

Wu, J., Liu, F., Sun, J., Wei, Q., Kang, W., Wang, F., Zhang, C., Zhao, M., Xu, S., & Han, B. (2024). Toxic effects of acaricide fenazaquin on development, hemolymph metabolome, and gut microbiome of honey bee (*Apis mellifera*) larvae. *Chemosphere*, *358*, 142207. https://doi.org/10.1016/j.chemosphere.2024.142207

Wu, X., Galbraith, D. A., Chatterjee, P., Jeong, H., Grozinger, C. M., & Yi, S. V. (2020). Lineage and parent-of-origin effects in DNA methylation of honey bees (*Apis mellifera*) revealed by reciprocal crosses and whole-genome bisulfite sequencing. *Genome Biology and Evolution*, *12*(8), 1482–1492. https://doi.org/10.1093/gbe/evaa133

Wu, Y., Zheng, Y., Chen, Y., Wang, S., Chen, Y., Hu, F., & Zheng, H. (2021). Honey bee (*Apis mellifera*) gut microbiota promotes host endogenous detoxification capability via regulation of P450 gene expression in the digestive tract. *Microbial Biotechnology*, *13*(4), 1201–1212. https://doi.org/10.1111/1751-7915.13579

Xi, B., Gu, H., Baniasadi, H., & Raftery, D. (2014). Statistical analysis and modeling of mass spectrometry-based metabolomics data. *Methods in Molecular Biology*, *1198*, 333–353. https://doi.org/10.1007/978-1-4939-1258-2_22

Xu, R., Ma, B., Yang, Y., Dong, X., Li, J., Xu, X., & Fang, Y. (2024). Proteome-metabolome profiling of wax gland complex reveals functional changes in honey bee, *Apis mellifera* L. *iScience*, *27*(3), 109279. https://doi.org/10.1016/j.isci.2024.109279

Yan, H., Simola, D. F., Bonasio, R., Liebig, J., Berger, S. L., & Reinberg, D. (2014). Eusocial insects as emerging models for behavioural epigenetics. *Nature Reviews. Genetics*, *15*(10), 677–688. https://doi.org/10.1038/nrg3787

Yan, S., Mu, G., Yuan, Y., Xu, H., Song, H., & Xue, X. (2024). Exploring the formation of chemical markers in chaste honey by comparative metabolomics: From nectar to mature honey. *Journal of Agricultural and Food Chemistry*, *72*(18), 10596–10604. https://doi.org/10.1021/acs.jafc.4c01340

Yong, W.-S., Hsu, F.-M., & Chen, P.-Y. (2016). Profiling genome-wide DNA methylation. *Epigenetics & Chromatin*, *9*(1), 26. https://doi.org/10.1186/s13072-016-0075-3

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208. https://doi.org/10.1038/ng1702

Yun, J.-H., Jung, M.-J., Kim, P. S., & Bae, J.-W. (2018). Social status shapes the bacterial and fungal gut communities of the honey bee. *Scientific Reports*, *8*(1), 2019. https://doi.org/10.1038/s41598-018-19860-7

Yusoff, Y. M., Abbott, G., Young, L., & Edrada-Ebel, R. (2022). Metabolomic profiling of Malaysian and New Zealand honey using concatenated NMR and HRMS datasets. *Metabolites*, *12*(1), 85. https://doi.org/10.3390/metabo12010085

Zhang, Z., Mu, X., Cao, Q., Shi, Y., Hu, X., & Zheng, H. (2022). Honey bee gut *Lactobacillus* modulates host learning and memory behaviors via regulating tryptophan metabolism. *Nature Communications*, *13*(1), 2037. https://doi.org/10.1038/s41467-022-29760-0

Zhang, Z., Mu, X., Shi, Y., & Zheng, H. (2022). Distinct roles of honey bee gut bacteria on host metabolism and neurological processes. *Microbiology Spectrum*, *10*(2), e0243821. https://doi.org/10.1128/spectrum.02438-21

Zhang, W., Wang, L., Zhao, Y., Wang, Y., Chen, C., Hu, Y., Zhu, Y., Sun, H., Cheng, Y., Sun, Q., Zhang, J., & Chen, D. (2022). Single-cell transcriptomic analysis of honey bee brains identifies vitellogenin as caste differentiation-related factor. *iScience*, *25*(7), 104643. https://doi.org/10.1016/j.isci.2022.104643

Zhang, Y., He, X. J., Barron, A. B., Li, Z., Jin, M. J., Wang, Z. L., Huang, Q., Zhang, L. Z., Wu, X. B., Yan, W. Y., & Zeng, Z. J. (2023). The diverging epigenomic landscapes of honey bee queens and workers revealed by multiomic sequencing. *Insect Biochemistry and Molecular Biology*, *155*, 103929. https://doi.org/10.1016/j.ibmb.2023.103929

Zhang, Y., Li, Z., He, X., Wang, Z., & Zeng, Z. (2023). H3K4me1 modification functions in caste differentiation in honey bees. *International Journal of Molecular Sciences*, *24*(7), 6217. https://doi.org/10.3390/ijms24076217

Zhao, H., Li, G., Guo, D., Wang, Y., Liu, Q., Gao, Z., Wang, H., Liu, Z., Guo, X., & Xu, B. (2020). Transcriptomic and metabolomic landscape of the molecular effects of glyphosate commercial formulation on *Apis mellifera ligustica* and *Apis cerana cerana*. *The Science of the Total Environment*, *744*, 140819. https://doi.org/10.1016/j.scitotenv.2020.140819

Zheng, H., Powell, J. E., Steele, M. I., Dietrich, C., & Moran, N. A. (2017). Honey bee gut microbiota promotes host weight gain via bacterial metabolism and hormonal signaling. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(18), 4775–4780. https://doi.org/10.1073/pnas.1701819114

Zheng, H., Steele, M. I., Leonard, S. P., Motta, E. V. S., & Moran, N. A. (2018). Honey bees as models for gut microbiota research. *Lab Animal*, *47*(11), 317–325. https://doi.org/10.1038/s41684-018-0173-x

Zhong, S., Pan, L., Wang, Z., & Zeng, Z. (2024). Revealing changes in ovarian and hemolymphatic metabolites using widely targeted metabolomics between newly emerged and laying queens of honey bee (*Apis mellifera*). *Insects*, *15*(4), 263. https://doi.org/10.3390/insects15040263

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821–824. https://doi.org/10.1038/ng.2310

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